Computational Studies of Tau Protein: Implications for the Pathogenesis and Treatment of Neurodegenerative Diseases

By

Austin Huang

B.S. Electrical Engineering and Computer Science
University of California, Berkeley, 2002

M.S. Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 2005

SUBMITTED TO THE HARVARD-MIT DIVISION OF HEALTH SCIENCES AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN ELECTRICAL AND BIOMEDICAL ENGINEERING AT THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY

[June]
MAY 2009

Signature of Author: _____
Harvard-MIT Division of Health Sciences and Technology
May 18, 2009

Certified by: _____
Collin M. Stultz
W. M. Keck Associate Professor of Biomedical Engineering
Associate Professor of Health Sciences and Technology
Thesis Supervisor

Accepted by: _____
Ram Sasisekharan,
PhD/Director, Harvard-MIT Division of Health Sciences and Technology/Edward Hood Taplin Professor of Health Sciences & Technology and Biological Engineering

Computational Studies of Tau Protein: Implications for the Pathogenesis and Treatment of Neurodegenerative Diseases

By

Austin Huang

Submitted to the Harvard-MIT Division of Health Sciences and Technology in partial fulfillment of the requirements for the degree of doctor of philosophy in electrical and biomedical engineering at the Massachusetts Institute of Technology

Abstract

Tau protein is the primary constituent of protein aggregates known as neurofibrillary tangles, a pathological hallmark of Alzheimer's disease (AD). Previous studies suggest that tau protein may play a contributing role in neurodegenerative diseases such as AD. Thus characterizing the structural properties of tau is critical to understanding disease pathogenesis. However, obtaining a detailed structural description of tau protein has been difficult because it belongs to a class of heteropolymers known as intrinsically disordered proteins (IDPs). Unlike most proteins, IDPs adopt many distinct conformations under physiological conditions. In spite of their disordered nature, evidence exists that such proteins may exhibit residual structural preferences. In this work, models of tau are constructed to characterize these structural preferences. We begin by performing molecular dynamics simulations to study the inherent conformational preferences of the minimal tau subsequence required for in vitro aggregation. To model residual structure in larger regions of tau, we developed a novel method called Energy-minima Mapping and Weighting (EMW). The method samples energetically favorable conformations within an IDP and uses these structures to construct ensembles that are consistent with experimental data. This method is tested on a region of another IDP, $p21^{Waf1/Cip1/Sdi1}$(145-164), for which crystal structures of substrate-bound conformations are available. Residual conformational preferences identified using EMW were found to be comparable to crystal structures from substrate-bound conformations of $p21^{Waf1/Cip1/Sdi1}$(145-164). By applying EMW to tau, we find disease-associated forms of tau exhibit a conformational preference for extended conformations near the aggregation-initiating region. Since an increased preference for extended states may facilitate the propagation of cross-β conformation associated with aggregated forms of tau, these results help to explain how local conformational preferences in disease-associated states can promote the formation of tau aggregates. Finally, we examine limitations of the current methods for characterizing IDPs such as tau and discuss future directions in the modeling of these proteins.

Chapter 1: Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder characterized by progressive memory loss, cognitive dysfunction, and behavioral disturbances [3]. The disease has a high prevalence, afflicting approximately 18 million people worldwide and is the most common cause of senile dementia [4]. The two pathological hallmarks of Alzheimer's disease are extracellular protein aggregates of amyloid-β (Aβ), known as amyloid plaques, and intracellular protein aggregates of tau protein, known as neurofibrillary tangles [5]. Much data suggests that the proteins which constitute these aggregates, Aβ and tau, also play a role in disease pathogenesis [6-10]. A structural description of these proteins is required to understand the conformational transitions accompanying aggregation into potentially toxic forms and to assist in the design of therapeutics targeting these proteins [11].

Despite that the majority of AD research has focused on Aβ, much evidence suggests that tau dysfunction contributes to disease progression in AD [6, 7, 12, 13]. Tau protein also plays an important role in a related family of neurodegenerative diseases, known as tauopathies, which are neurodegenerative disorders characterized by pathological aggregation of tau [14]. Tau protein belongs to a class of heteropolymers known as intrinsically disordered proteins (IDPs) [11]. These proteins are sometimes referred to as natively unfolded proteins (NUPs) or intrinsically unstructured proteins (IUPs). In contrast to most proteins, which fold into a unique, three-dimensional structure or at least contain large regions of structure, IDPs fluctuate between many distinct conformations under physiological conditions. Presently there are no existing experimental methods to fully characterize the set of structures populated by these proteins.

In this work, we combine biophysical modeling and conformational sampling approaches with published experimental measurements to characterize the structural properties of tau protein. Thus, one can obtain detailed structural insights that are not available from experiments alone. The thesis is organized as follows:

Chapter 2 provides an overview of recent experimental and modeling approaches for characterizing structural properties of IDPs involved in neurodegenerative diseases.

Chapter 3 discusses molecular dynamics simulations performed on a peptide corresponding to a key tau subsequence which is required for aggregation in vitro. There were two motivations for this study. First, conformational preferences of this subsequence were of inherent interest due to its importance in tau aggregation. Second, these simulations were used to evaluate implicit solvent models for the purpose of sampling conformational minima. Using these methods, we find that the aggregation-initiating sequence has an intrinsic propensity for extended conformations. Furthermore, we identified an implicit solvent potential for efficient sampling of conformational minima.

In order to identify residual conformational preferences in larger regions of tau, we developed a novel semi-empirical method for constructing conformational ensembles of intrinsically disordered proteins. This protocol is discussed and used in the studies described in chapter 4 and the appendix. We initially tested the method on a region of the intrinsically disordered protein $p21^{Waf1/Cip1/Sdi1}$ (appendix). Unlike tau, $p21^{Waf1/Cip1/Sdi1}$ is an intrinsically disordered protein for which crystal and NMR structures of substrate-bound subsequences exist [15, 16]. Thus, we compared structural properties described by our method against known bound conformations. This analysis showed that local conformational preferences in the unfolded state of $p21^{Waf1/Cip1/Sdi1}$ identified by our method were comparable to structured substrate-bound

conformations. This work was performed with Veena Venkatachalam, an undergraduate student, and in collaboration with James Chou [17]. The method was then applied to both wild-type and disease-associated mutant forms of tau protein (Chapter 4). The resulting conformational ensembles describe how changes in local conformational preferences between normal and disease-associated forms of tau can contribute to differences in their propensity to aggregate.

Recently, additional structural data for the microtubule-binding repeat domain of tau (referred to as K18) have been published [18, 19]. Chapter 5 discusses work to incorporate these data into models of K18. In addition, we attempted to test an alternate modeling approach which does not require fitting to experimental data. This model is based on the hypothesis that conformational preferences of sequentially long-range positions can be approximated as being independent. In Chapter 6, we discuss limitations of current experimental and modeling approaches to characterizing structure in tau (and IDPs in general) and future directions for research.

Chapter 2: Finding Order within Disorder - Elucidating the Structure of Proteins Associated with Neurodegenerative Disease

(This work was published as A. Huang and Stultz CM., "Finding Order within Disorder - Elucidating the Structure of Proteins Associated with Neurodegenerative Disease," Future Medicinal Chemistry (accepted), 2009.)

Abstract

A number of neurodegenerative disorders such as Alzheimer's disease and Parkinson's disease involve the formation of protein aggregates. The primary constituent of these aggregates belongs to a unique class of heteropolymers called intrinsically disordered proteins (IDPs). While many proteins fold to a unique conformation that is determined by their amino acid sequence, IDPs do not adopt a single well-defined conformation in solution. Instead they populate a heterogeneous set of conformers under physiological conditions. Interestingly, despite this intrinsic propensity for disorder a number of these proteins can form ordered aggregates both *in vitro* and *in vivo*. As the formation of these relatively ordered aggregates may play an important role in disease pathogenesis, a detailed structural characterization of these proteins and their mechanism of aggregation is of critical importance. However, given their inherent complexity and heterogeneity, new methods are needed to decode the diversity of structures that make up the unfolded ensemble of these systems. Here we discuss recent advances in the structural analysis and modeling of IDPs involved in neurodegenerative diseases, and outline future directions in the development of therapies that are designed to prevent their aggregation.

Introduction

Many proteins encoded by the human genome fluctuate about a well-defined three dimensional structure during their biological lifetimes. This observation has lead to the often stated, and amply validated, structure-function paradigm; i.e., a protein's function is determined by its three dimensional structure [20]. However, it is increasingly apparent that a number of proteins in the human proteome do not adopt well defined three-dimensional structures under physiologic conditions [21-23]. These intrinsically disordered proteins (IDPs) stand in stark contrast to the archetypal structure-function paradigm and therefore represent unique and interesting biological heteropolymers whose study may lead to insights into the relationship between amino-acid sequence and structure. More importantly, deciphering the relationship between structure and function for these systems is not purely an academic exercise. Many neurodegenerative diseases have been associated with abnormal/excessive aggregation of IDPs. Alzheimer's Disease, one of the most prevalent forms of Dementia in the US, is associated with two types of IDP aggregates. Extracellular aggregates known as senile plaques are composed of amyloid-β peptide, a cleavage product of amyloid-precursor protein (APP) (Figure 1, APP) and intraneuronal aggregates, known as neurofibrillary tangles (NFTs), are composed of the IDP tau protein (Figure 1, tau) [24, 25]. Parkinson's Disease associated Dementia (PDD) or Dementia with Lewy Body Disease (DLB) is the second most common form of dementia and is characterized by aggregates, known as Lewy bodies, which are primarily composed of the intrinsically disordered protein α-synuclein (Figure 1, α-synuclein) [26].

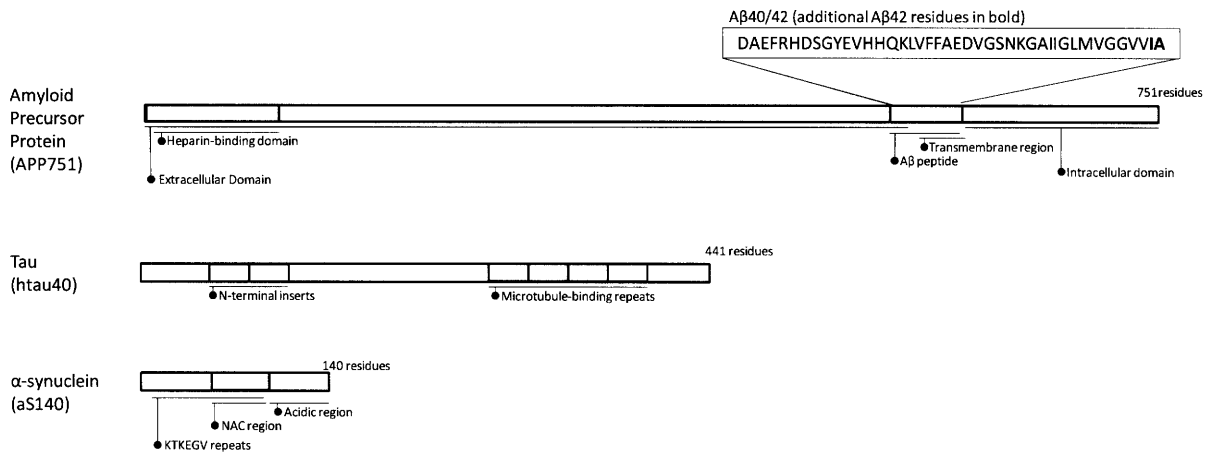


Figure 1 Primary structure of the common isoforms of proteins involved in neurodegenerative diseases. Amyloid Precursor Protein contains a 40-42 residue segment (Aβ40/42) that is found in senile plaques and aggregates of tau protein are found in Neurofibrillary tangles. Both types of aggregates are found in patients with Alzheimer's Disease. Aggregates of α-synuclein are found in patients with patients with Parkinson's Disease and Dementia with Lewy Body Disease.

Substantial evidence exists to suggest that these IDPs play an important role in these neurodegenerative diseases. First, several studies argue that mutations in APP, tau, and α-synuclein result in hereditary forms of neurodegenerative diseases [27-31]. In animal and cell models, mutant forms of APP and tau, or overexpression of these proteins, results in disease phenotypes [6-8, 32, 33]. Neurofibrillary tangles as well as levels of soluble Aβ oligomers, are correlated with progression of AD and promisingly, disruption of aggregation has been found to prevent or reverse disease progression in cell and animal models [6, 7, 9, 12, 34, 35]. Taken together, these data suggest that pathological aggregation of these intrinsically disordered proteins may be a key contributor to these disease processes.

IDPs typically fluctuate between different conformations under physiologic conditions, leading to an ensemble of structurally dissimilar states. Recent studies, however, suggest that the aggregation process of a number of IDPs, like Aβ, tau and α-synuclein, involves the formation of partially folded intermediates that self associate to form complexes that contain considerable

cross β-structure [36, 37]. Hence the aggregation process for many of these IDPs is quite complex. Given the likely importance of aggregation to neuronal degeneration and dysfunction, understanding the nature of the unfolded state of these proteins and their ability to form ordered aggregates is an important fundamental problem in biology and medicine. Furthermore, a comprehensive understanding of the aggregation mechanism of these heteropolymers may lead to novel therapies that prevent their aggregation.

Typically, structural information on proteins is obtained using well-established techniques. In order to solve a protein's structure with x-ray crystallography (the most commonly used method for structure determination), the protein must first be crystallized so that a meaningful diffraction pattern can be obtained. Crystallization is successful when the different protein molecules in the crystal have nearly identical structures and orientations. In the case of IDPs however, the inherent structural heterogeneity of the system makes successful crystallization not possible. Clearly protein x-ray crystallography, which effectively determines the average of the different structures in the protein crystal, is not an appropriate technique to study the structure of IDPs, which are inherently heterogeneous. Any useful characterization of the structure of these proteins must therefore capture the inherent diversity of structures that populate the unfolded ensemble. It is in this regard that physically based models can play an important role. Computer simulations of proteins, for example, have shed considerable light on the aggregation mechanism of polygutamine containing proteins, which are intrinsically disordered [38, 39]. Moreover, approaches that combine experimental results with computational methods can be used to construct detailed structural models of the unfolded state of IDPs. The resulting insights contribute to our understanding of the aggregation process and may ultimately be useful for designing compounds that prevent IDP aggregation.

In this chapter we discuss methods and strategies for constructing detailed models of the unfolded state of IDPs that are believed to play a role in neurodegenerative disorders. We illustrate how such information has shed light on the pathogenesis of aggregation and suggest how these insights can be used to initiate new drug design strategies.

Characterizing the Structure of an Intrinsically Disordered Protein

The early work of Anfinsen coupled with the advent of protein crystallography helped to establish the paradigm that each amino-acid sequence is associated with a unique three dimensional structure [40-42]. In this context, structural characterizations of proteins have an unambiguous meaning; i.e., characterization of a protein's structure involves finding this unique conformation. As more IDPs have been discovered this classic paradigm must be recast in a form that is appropriate for systems that do not adopt a well-defined structure in solution. In a sense the central question for these systems becomes: what does it mean to characterize the structure of an IDP?

Early views of the unfolded state of proteins described disordered polypeptides as random coils devoid of structural preferences [43]. However, recent reviews of molecular volume characteristics and other properties of intrinsically disordered proteins have made it clear that there exists a spectrum of disorder and that a generic random coil is not always an adequate description for the structural properties of IDPS [21-23]. Molecular volume scaling properties range from highly disordered random-coil like conformations to relatively collapsed premolten globule like states [22, 23]. In studying these proteins, several questions commonly arise. Is a featureless random coil adequate to explain the experimental measurements? Are there global

conformations that are strongly preferred? Are there local conformational preferences? Most importantly - how are conformational preferences related to protein function, aggregation, and disease pathogenesis?

Insights into the physical basis of the structural heterogeneity within IDPs can be garnered from a review of the protein folding literature. It has been suggested that proteins that adopt a well-defined structure in solution fold on "funnel shaped" energy surfaces [44, 45]. The folding-funnel hypothesis postulates that in spite of the astronomical number of conformations associated with the unfolded state [46], the conformational free energy surface is shaped like a funnel and this drives the protein towards its folded conformation (Figure 2A). Although there are many different possible conformations for the protein, the shape of the surface ensures that the protein reliably folds to a unique region of conformational space, corresponding to the folded state [45, 47]. Unlike the conformational free energy landscape of a natively folded protein, the conformational free energy landscape of an intrinsically disordered protein lacks a prominent folding funnel (Figure 2B) [23]. As such, the protein can adopt multiple low energy conformations on the energy surface, corresponding to local energy minima. For folded proteins, characterization of the structure entails finding the low energy conformer at the basin of the funnel (Figure 2A). By contrast, characterization of the unfolded state of an IDP necessarily encompasses an enumeration of its accessible low energy conformations.

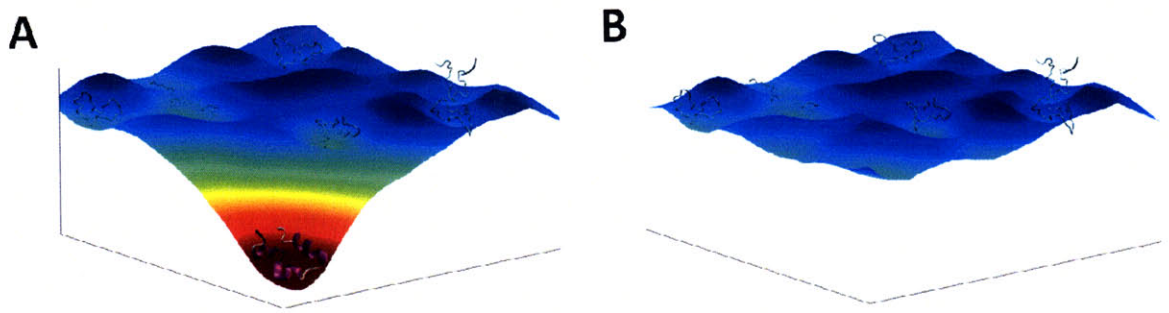


Figure 2 Schematic of a conformational free energy landscapes. A) The conformational free-energy landscape of a folded protein exhibits a deep well-defined minimum corresponding to the native conformation. B) The conformational free-energy landscape of an intrinsically disordered protein lacks a deep free-energy minimum. Thus the "native state" consists of an ensemble of interconverting conformations.

It is important to note that enumerating accessible low energy states is not equivalent to enumerating all possible conformations. For many systems the most prevalent structures span a range that may exhibit strong conformational preferences. In the case of tau protein, for example, there are data to suggest that some local structural motifs are strongly preferred over others [19, 48, 49]. Studies such as these suggest that the unfolded state of IDPs have their own taxonomy and therefore may exhibit distinct preferences for particular structural states. Consequently, a comprehensive characterization of the unfolded state of an IDP should include a description of these conformational preferences.

It is also noteworthy that not all structures within a given unfolded ensemble are created equal. There are data to suggest that for many systems aggregation proceeds along a nucleation growth mechanism that is facilitated by the formation of partially folded intermediates [36, 50]. Structures within the unfolded ensemble that have characteristics similar to these partially folded intermediates may facilitate the aggregation process *in vitro* (Figure 3). For such systems a central problem in the characterization of the unfolded state is the identification of such aggregation prone conformers. Once identified and isolated, one can design molecules that specifically prevent the self association of these problematic structures.

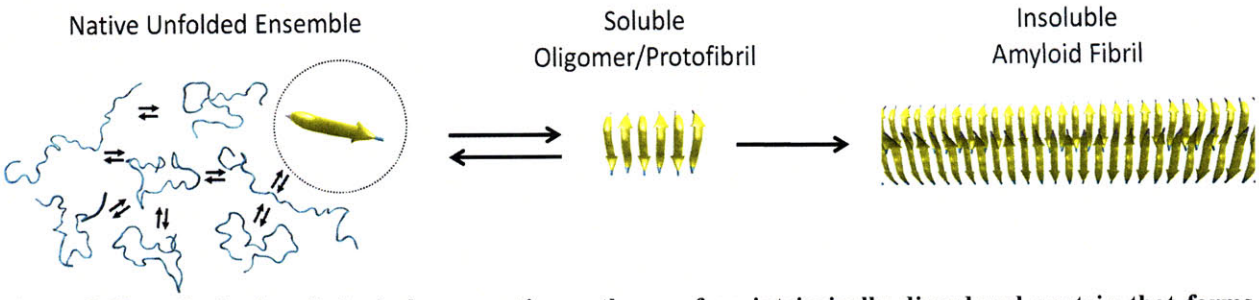


Figure 3 Hypothesized pathological aggregation pathway of an intrinsically disordered protein that forms amyloid fibrils. The circled structure represents an aggregation-prone conformer within the native unfolded ensemble. Conformers in the unfolded ensemble can act as partially folded intermediates, which are aggregation-prone. These aggregation-prone conformers initiate formation of soluble oligomers. Oligomers are extended and stabilized by the addition of monomers to form an insoluble amyloid fibril. We note that the structure of soluble oligomers are not known, therefore the above mechanism remains speculative.

Current Experimental Approaches to Studying Intrinsically Disordered Proteins

The experimental characterization of intrinsically disordered proteins is a relatively nascent field. Several excellent reviews that summarize current experimental methods for studying intrinsically disordered proteins have recently been published [50-53]. A number of methods have been fruitfully applied to gain insights into the unfolded ensemble of IDPs. While a vast and diverse array of methods have been utilized, including optical spectroscopy, small-angle X-ray scattering (SAXS), and dynamic light scattering, the most popular methods have been based on nuclear magnetic resonance (NMR) spectroscopy [51, 54]. [55]. While much of the remainder of this section focuses on NMR-based measures, we note that many other techniques provide complementary information that can yield important insights into the nature of the unfolded state. For example, optical spectroscopic methods can greatly complement NMR-derived measurements, as the timescales and experimental conditions of such optical spectroscopic methods are different from NMR-based methods. In fact, initial determination of whether a protein is intrinsically disordered is frequently accomplished using circular dichroism (CD), an

optical spectroscopic method [55]. A more complete discussion of the relative strengths of other methods can be found in the aforementioned references [50-53].

A number of different NMR-based measures have been used to glean information about the unfolded state of IDPs. These methods include an array of nuclear magnetic resonance (NMR) techniques. For example, Nuclear Overhauser Effect (NOE) measurements are used to determine short-range contacts preserved in the unfolded ensemble and chemical shifts, residual dipolar couplings (RDCs) and ^{3}J couplings can provide residue-specific information regarding local conformational preferences. Methods such as pulse-field gradient diffusion – in addition to non-NMR techniques such as SAXS – are used as standard measures for comparing and contrasting the protein of interest with the standard random coil models.

The existence of multiple experimental methods that can be applied to IDPs has enabled important insights regarding the structural characteristics of these systems [52]. Nevertheless, it is important to realize that most experimental measurements represent ensemble averages over a large number of distinct conformations. Therefore it is often difficult to draw detailed conclusions about the types of conformers that populate the unfolded ensemble from these data alone. Even so, it is becoming clear that some experimental metrics are much more useful than others. NOEs can be quite useful in the analysis of protein structure because they yield precise distance measures between residues that might be separated in the amino-acid sequence [56]. Yet, given the relatively short internuclear distance that is explored by NOE experiments, only residues that remain in very close contact in the majority of structures in the ensemble will give reliable NOE data. In particular, NOEs that are indicative of long range contacts are usually not observed in IDPs. By contrast, paramagnetic relaxation experiments can yield distance measures

that are significantly larger and therefore these data may be quite useful for identifying long range contacts that are preferentially populated in an unfolded ensemble [57].

Recent data suggest that RDC measurements may also be particularly useful for understanding the nature of the unfolded state. Indeed, it has been argued that RDCs encode information about the average global structure of the protein [58, 59]. The greater information content of RDCs arises from the fact that prior to measuring RDCs, proteins are placed in an alignment medium, where the exact alignment of the protein is determined by the protein's global structure [59, 60].

More recent non-NMR based experimental methods have employed disulfide scrambling assays to capture particular conformers within the unfolded state [61]. The underlying idea is outlined in Figure 4. In this example, cysteine residues are introduced into an IDP at two positions. In Figure 4 the protein is depicted as being able to adopt three distinct conformations, C_1, C_2 and C_3. The introduction of cysteine residues in the sequence is done to facilitate disulfide bond formation in state C_3, but not in conformers C_1 and C_2. Hence under oxidizing conditions a disulfide bonded form of conformer C_3 (conformer C_4) is formed. Since the formation of structure C_4 is essentially irreversible, the equilibrium becomes shifted towards this "trapped" state by the laws mass action. Therefore, if we allow the redox reaction to proceed for a significant period of time, a solution that is enriched in conformer C_4 will be obtained. This approach, which has been pioneered by Chang, has been applied to α-synuclein [61, 62]. As some of the trapped disulfide-bonded isomers had an increased propensity to aggregate relative to the wild-type protein, these data argue that the unfolded state of α-synuclein in solution contains a variety of different conformations in solution, where some conformations are more aggregation-prone than others. One drawback of their approach is that the authors had no model

for the unfolded state, the residues to mutate into cysteine were not chosen based on structural considerations; e.g., residues known to be phosphorylated in wild-type α-synuclein were mutated [61]. Clearly a more detailed model of the unfolded state would help to effectively identify aggregation prone conformers.

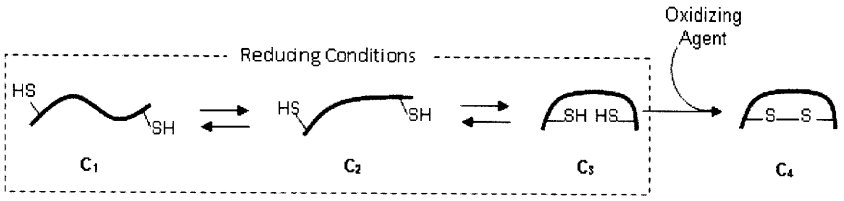


Figure 4 Trapping of specific conformations within an unfolded ensemble

One challenge facing many of these experimental methods is the sensitivity of the unfolded ensemble to perturbations in experimental conditions. Experimental conditions often need to deviate from physiological conditions to optimize the signal to noise ratio for a given experimental measurement. Varying experimental conditions to study folded proteins is widely employed and generally well founded. Most notably, crystallization conditions for folded proteins may deviate substantially from physiological conditions, yet the solved structures have been shown to be physiologically relevant in a many cases (e.g., [63, 64]). However, the case may be quite different for an IDP. In the absence of a deep native basin at the bottom of a folding funnel on the free energy landscape, it is unclear to what extent the distribution of conformers will be perturbed by even minor changes to the protein's chemical environment. Thus, independent measurements made on the unfolded ensemble may reflect different ensembles and these ensembles may also differ from the conformational ensemble under physiological conditions. For example the conformational equilibrium of α-synuclein is substantially changed by modifying buffer conditions [65, 66].

In short, the relevance of *in vitro* studies to the structure of these proteins *in vivo* remains an open question. Nevertheless, methods that use data obtained from *in vitro* studies to guide the development of unfolded ensembles provide a rigorous framework to develop and test methods for modeling the unfolded state. Moreover, these studies lead to testable hypotheses that can be extended to *in vivo* studies.

Methods for Constructing Models of the Unfolded Ensemble

Intuitions derived from an analysis of experimental measurements made on folded proteins can lead to misinterpretation when these same methods are applied to an intrinsically disordered protein. Model based approaches that incorporate experimental measurements to aid in the analysis of experimental observations allow one to interpret experimental findings within the context of structural ensembles that describe the types of conformers that populate the unfolded ensemble. As such, these methods have greatly improved our understanding of the unfolded states of proteins and have clarified our understanding of experimental data.

One early method used to model the unfolded state of biological heteropolymers is the statistical coil model. These models are called statistical coil models because their ensembles are defined by a set of random coil structures sampled from residue-specific statistical distributions. They typically account for a very limited set of factors in the unfolded state; e.g. residue-specific backbone dihedral angle propensities that are parameterized based on loop regions from the protein data bank and steric clashes [67, 68]. From these distributions a conformational ensemble for any IDP can be created by sampling conformations for each residue. These models have been used to interpret measurements made on a number of IDPs such as tau protein and in many

instances qualitative agreement with experimental measurements has been obtained [19, 67, 68]. Nevertheless, the simplicity of the statistical coil model is both a strength and a weakness. In general, the residue-specific conformational distributions fully parameterize the model - there is no additional system-specific parameter fitting required. At the same time it is difficult to gain system-specific insight from the model because conformational correlations between distant residues in the sequence are not captured. In addition, it has been shown that the aggregation properties of some IDPs can be dramatically altered by introducing single residue changes in the sequence (e.g., [69-72]). NMR based measurements for such mutant sequences exhibit very small changes relative to the wild-type protein [70, 72]. In these cases it is not clear that these statistical models, which typically have qualitative agreement with experiment, is sufficient to capture ensemble differences between wildtype and mutant proteins that correlate with distinct aggregation properties. Consequently, other methods that yield more quantitative agreement with experiment are needed to decipher the relationship between small changes in the experimentally determined quantities and aggregation behavior.

Molecular dynamics simulations have been employed to generate low energy structures that may represent the unfolded state ensemble. Extended-ensemble algorithms such as replica exchange can aid in efficiently sampling conformational transitions yielding a heterogenous set of structures [73-75]. In some instances (as we discuss in the next section), these methods have yielding important insights into the types of structures that populate the unfolded state. However, inaccuracies in the underlying potential function and the approximate nature of the solvent model can make reliable agreement with experiment difficult.

Recently a number of approaches have been developed to create models that yield improved agreement experiment [76]. While they differ in their details, most of them have similar

overarching themes. Many begin by first creating a library of protein conformations using a conformational sampling algorithm (e.g., Monte Carlo, Molecular Dynamics, etc.) which can then interpreted as the ensemble. In some cases, the resulting set of sampled conformations is then used to generate subsets that represent the unfolded ensemble. In addition to a set of conformers in the unfolded ensemble, some methods also associate a weight with each conformer. A conformer's weight represents the probability that the protein in question adopts that specific conformation. Sampling and initial weight assignment is then followed by additional model optimization to fit the experimental data. Examples of these approaches are the ENSEMBLE algorithm [77, 78], sample-and-select (SAS) [79], ensemble optimization method (EOM) [80], and energy-minima mapping and weighting (EMW) method [49]. In each case, construction of the ensemble is formulated as an optimization problem where the function to be minimized captures the difference between ensemble-averaged structural quantities of the model and the measured experimental observables.

As the structural ensemble is explicitly represented as a discrete set of structures, this type of model can predict any experimental quantity as long as there is a function $S(X_i)$ that maps each conformation, X_i, in the ensemble to an experimental measurement. For example, $S(X_i)$ can be the chemical shift or RDC of a given residue in the protein and therefore obtained using a number of established algorithms, which calculate chemical shifts and residual dipolar couplings associated with a particular structure [2, 81, 82]. Once the experimental value in question can be calculated for a given conformer in the protein, the ensemble average predicted value can be expressed as a sum of the contributions due to each conformer:

$$S_E(w_i, X_i) = \sum_{i=1}^{N} w_i S(X_i) \tag{2.1}$$

where the weight, w_i, is the probability that the protein adopts conformation X_i, and $S_E(w_i, X_i)$ is the ensemble averaged value. The summation goes over the set of N structures in the ensemble. Ensemble members are chosen to minimize the error between the calculated ensemble averaged value and the corresponding experimentally determined value, which is typically obtained on a solution containing the IDP of interest. In algorithms such as SAS and EOM in which each structure is given equal weight and structures are sampled from a conformational library, this can expressed as constraining all values of w_j to evaluate to 1/N or 0, where N is the number of selected conformers. Furthermore, the requirement that a limited number of conformers is used to fit the experimental data, can be expressed by limiting the number of weights, w_j, that are nonzero. Several approaches have been utilized in the optimization of this objective function, including simulated annealing and genetic algorithms [79, 80].

There are two primary difficulties with these approaches. First, the conformational library must encompass the most prevalent conformations. As ensemble members are chosen from the conformational library, it will not be possible to obtain a physiologically relevant solution if the originally library does not contain conformers that are present with high probability in the unfolded ensemble. The second difficulty is the degeneracy of the solution space. Given that accurate modeling of an unfolded protein is an undetermined problem, it is likely that there are a number of different ensembles that agree with any given set of experimental data. By making additional independent measurements, one can reduce the size solution space by many orders of magnitude, as was shown by Choy et al for a small system [77]. As yet, however, a comparable analysis has not yet been performed for a full protein, which has a substantially larger number of degrees of freedom. Another approach to addressing the problem of degeneracy is to generate

multiple solutions that are consistent with a given set of experimental data and to find characteristics common to all solutions [49]. In other words, given the underdetermined nature of the problem, it is not clear how to determine when one has the "correct" ensemble. However, structural motifs that consistently appear in all independent ensembles are likely to also be present in the "correct" ensemble.

The application of these methods has lead to important insights on the nature of the unfolded state of several IDPs and their mechanism of aggregation. Below we highlight how these methods have been applied to distinct IDPs.

Modeling IDPs Associated with Neurodegenerative Disorders

Amyloid-β

Alzheimer's disease (AD) is the most common form of senile dementia, characterized by memory loss, personality changes, and global cognitive dysfunction [10]. AD is associated with two pathological hallmarks - extracellular amyloid plaques and intracellular neurofibrillary tangles composed primarily of amyloid-β peptide (Aβ) and hyperphosphorylated tau protein, respectively [14, 83, 84]. Both Aβ and tau are intrinsically disordered proteins. The IDP Aβ is a 39 to 43 residue peptide produced by cleavage of amyloid precursor protein (APP) by β-secretase and γ-secretase [36, 85] (Figure 1).

Despite the fact that Aβ is intrinsically disordered, it can be made to adopt a relatively restricted set of conformers under the right experimental conditions. For example, structural ensembles were constructed from NOE data obtained on Aβ(1-40) in an aqueous sodium dodecyl

sulfate (SDS) environment [86]. APP is a membrane-bound protein and Aβ is believed to include membrane-spanning residues of APP; thus the micelle-bound structure may reflect the initial conformation of Aβ immediately after cleavage from APP. While these data suggest that that Aβ(1-40) contains residual helical structure in a membrane environment, they are not likely to be representative of the aqueous monomeric structure which aggregates into amyloid fibrils. Structures derived from Aβ(10-35) in solution are a better mimic the physiologic environment of monomeric Aβ, though the exclusion of terminal residues may affect the conformational ensemble [87]. NMR data on Aβ(10-35) in solution have been obtained and suggest that the structure of peptide in solution is relatively compact and devoid of any secondary structure [87]. The structure also contains notable hydrophobic clusters at its core and also on its surface. In a later work, Hou et al measured ^{1}H, ^{15}N, and ^{13}C chemical shifts for Aβ40 and Aβ42 [88]. This study identified local conformational preferences for β-strand structure in hydrophobic regions (17-21 and 31-36) and turn conformations (7-11 and 20-26). While such studies shed light on potential conformers of Aβ or ensemble averaged characteristics, they also may not fully capture the conformational heterogeneity of the disordered protein.

Due to its relatively small size, molecular modeling studies of Aβ can be performed rather efficiently, yielding insights into the structure of, and transitions between its monomeric and aggregated forms. Extensive molecular dynamics simulations of a truncated form of Aβ (residues 10-35) were performed to obtain a conformational ensemble of the peptide in solution [89]. The model ensemble was compared with an NMR structure of Aβ(10-35) in solution that was obtained from NOE measurements. Quite remarkably, the ensemble obtained by simulation was able to reproduce the majority of the NOE constraints used for Aβ(10-35). More importantly, the ensemble derived from the simulations was structurally more diverse than the original

ensemble modeled from the NOE constraints alone. These results exemplify the difficulty of interpreting NOE-derived model ensembles of intrinsically disordered proteins as there may exist a number of different ensembles that agree with a given set of experimental data.

Several studies have examined the transition of Aβ monomers into aggregation-inducing intermediates. Since these studies suggest that cleavage of APP into Aβ occurs in early endosomes, Khandogin and Brooks used constant pH molecular dynamics (CPHMD) to probe the effect of endosomic pH on the conformational ensemble of Aβ [90]. These simulations found a pH dependent transition of the central hydrophobic residues of Aβ from helical to β-turn conformations, and suggest that the conformational ensemble at endosomal pH includes exposed hydrophobic residues and β-structure, properties consistent with aggregation initiation. These findings support the notion that endosomal pH may play a role in facilitating pathological conformational transitions of Aβ. Xu et al examined the conformational transition of Aβ from a helical membrane-bound form into an aqueous β-sheet rich intermediate and identified four glycine residues critical to this conformational transition.

Numerous coarse-grained and all-atom molecular dynamics simulations have also been used to explore the transition of Aβ from monomers to oligomers and examine the stability of aggregate species [91-96]. Fawzi et al utilized coarse-grained molecular dynamics simulations, showing the different effects of disease-associated mutants on the structure and stability of Aβ(1-40) protofibrils [93]. All-atom simulations performed on a critical region of Aβ (residues 16-22) to examine oligomer growth show that small oligomers undergo a substantial amount of rearrangements to accommodate the addition of a monomer and that addition of a monomer occurs during a two-phase mechanism consisting of fast association, followed by a slow conformational rearrangement [92]. The study also suggests that oligomer extension is distinct

from fibril extension, as fibril extension involves far smaller conformational changes to the aggregate species. Taken together, these studies provide a detailed view of oligomerization as a distinct, more dynamic process than fibril extension. Such a distinction is critically important, as there are data to suggest that oligomers, and not fibrils, may be the critical toxic species in AD [9].

α-synuclein

Parkinson's disease (PD) is characterized by an involuntary tremor, muscle stiffness, bradykinesia (slow movement), and postural instability [97]. Many patients with PD also have cognitive dysfunction that can be a major cause of morbidity and mortality. These patients can be characterized as having either Parkinson's Disease Dementia (PDD) or Dementia with Lewy Body Disease (DLB) [98]. In both cases, the pathological hallmark consists of intraneuronal protein aggregates known as Lewy Bodies [97]. As with AD, familial forms of PDD or DLB have been helpful in providing genetic data to implicate proteins that may have a causative role in the disease. A key gene associated with familial forms of Parkinson's disease encodes the α-synuclein protein, an intrinsically disordered protein, which is also the primary protein species present in Lewy bodies [99]. α-synuclein appears in 3 isoforms but is predominantly found in a 140 residue isoform and is primarily expressed in neural tissue [100]. α-synuclein contains a central portion – the non-amyloid component (NAC) region – which is believed to play an important role in the formation of protein aggregates [101, 102]. Interestingly, the NAC region fragment is also a secondary constituent of amyloid plaques in Alzheimer's disease [103, 104].

Despite its intrinsic disorder, α-synuclein self-associates to form fibrils that contain considerable cross β-structure [37, 105, 106]. Several studies suggest that fibrillization involves a nucleation growth mechanism that involves interactions between ordered segments in the

protein [70, 107]. Spin labeling and EPR studies on α-synuclein fibrils suggests that the central portion of the molecule, which contains the NAC(8-18) fragment, is folded into a core that contains significant β-structure, while the C-terminus of the molecule is disordered and the N-terminus is structurally heterogeneous [108, 109]. The importance of β-structure in the aggregation process is supported by the observation that mutations which decrease the propensity for β-structure can reduce the ability of α-synuclein to aggregate *in vitro* [110]. While these observations have advanced our understanding of the aggregation process, they do not directly provide information about the types of conformers that populate the unfolded ensemble of monomeric α-synuclein in solution.

Recent studies using atomic force microscopy on α-synuclein monomers confirm that unfolded ensemble contains different classes of structures where some conformers have a relative abundance of β-like structure – a structural motif thought to play a role in the aggregation mechanism [66]. Further insights into the unfolded ensemble and the aggregation process have been obtained from studies on α-synuclein mutants. As previously discussed, a recent work engineered cysteine mutations into α-synuclein and subsequently used a disulfide scrambling assay to capture isomers with particular disulfide bonding patterns [61]. Since the authors had no model for the unfolded state, the residues to mutate were not chosen based structural considerations. For example, residues known to be phosphorylated in wild-type α-synuclein were mutated [61]. As some of the observed isomers had an increased propensity to aggregate relative to the wild-type protein, these data argue that α-synuclein contains a variety of different conformations in solution, where some conformations are more aggregation-prone than others. Clearly a more comprehensive understanding of the mechanism underlying α-synuclein

aggregation and fibril formation requires detailed knowledge of the conformations that populate the unfolded state.

A recent study combined data from spin-label NMR experiments and restrained molecular dynamics simulations to model the unfolded state of α-synuclein [57]. Distance restraints arising from paramagnetic relaxation enhancement experiments were incorporated into standard MD simulations to develop a model of the unfolded ensemble. These data suggested that α-synuclein is more compact than would be expected using a standard random coil statistics – a finding consistent with previous experimental observations [107]. More importantly, it was demonstrated that this compactness was driven by long range contacts in the protein [57]. A subsequent study used a statistical coil model to generate a model for the unfolded ensemble. Those results were then compared with RDCs from α-synuclein [58]. While qualitative agreement with experiment was obtained for the central region of the protein, relatively poor agreement was noted for residues near the N and C-termini. It was demonstrated that significant improvement could be obtained by including long range contacts between residues near the N-terminus and the C-terminus – a finding consistent with the notion that that long range contacts exist in the unfolded state [58]. Interestingly, data obtained from paramagnetic relaxation measurements made in the presence of substances known to promote aggregation in vitro, suggest that these long range interactions are released under aggregation-promoting conditions [111]. It has therefore been suggested that the release of long range contacts leads to the exposure of hydrophobic regions that can then facilitate self-association.

Tau Protein

Tau protein naturally occurs in six isoforms (the largest of which is 441 residues in length) and belongs to the family of microtubule-associated proteins. The C-terminal region of tau

contains four imperfect microtubule binding repeats (Figure 2) [112]. As many disease-associated mutations are found in this domain, it is thought to play an important role in the misfolding and aggregation of tau into neurofibrillary tangles [113]. Alongside AD there is an entire class of diseases, known as tauopathies, which are neurodegenerative disorders characterized by pathological aggregation of tau [14].

There are many unanswered questions regarding the mechanism underlying the formation of tau aggregates. One issue is the role of phosphorylated tau in the formation of NFTs. Aggregated tau typically exists in a highly phosphorylated form [114]. Although nineteen different phosphorylation sites on tau have been identified, little is known about the precise role that phosphorylated isoforms have on the aggregatory process and cellular death [115]. There are data to suggest that phosphorylation at specific sites can affect the kinetics and thermodynamics of tau-microtubule binding, resulting in a decreased affinity of tau for microtubules [116]. In addition, mutations of tau that mimic the effects of phosphorylation at specific sites can enhance the formation of tau fibrils *in vitro* [117, 118]. These data suggest that phosphorylation may directly play a role in tau aggregation *in vivo* and *in vitro*. Methods that assess the effect of phosphorylation on the structure of tau may help to decipher the complex role that phosphorylation plays in tau pathology. In addition to hyperphosphorylation, a number of missense mutations in tau have been linked to neurodegenerative disorders that bear the pathological hallmark of increased NFT formation [119]. While many of these mutations have multifaceted effects on tau expression, a number of them can also promote tau aggregation *in vitro* [120, 121]. Consequently, deriving methods that can model the unfolded state of tau may help to decode the role that structural changes in the unfolded ensemble play in the development of NFTs.

Tau protein was initially identified as an intrinsically disordered protein through the use of CD measurements, SAXS, and Fourier Transform Infrared Spectroscopy (FTIR) [55]. Analysis of secondary chemical shifts and modeling studies of WT and disease-associated mutant forms of the tau suggest local structural preferences such as extended structure in aggregation-initiating regions of the microtubule-binding-repeat domain [48, 49, 122-124]. These experiments provide insight into the gross average structural properties of tau. However, more detailed models of the unfolded state are needed to obtain a more complete picture of structural propensities within the unfolded ensemble.

RDC measurements were made on constructs representing the microtubule-binding repeat domains of tau and a statistical coil model was used to interpret the RDC measurements [19]. However, the model was unable to fully reproduce the experimental data and additional molecular dynamics simulations were used to re-parameterize the residue-dependent distributions, yielding calculated RDC values that were in agreement with experiment. The resulting structures suggested that short stretches in each repeat domain adopted turn conformations with relatively high frequency. These data demonstrate that the coupling of NMR data and molecular simulations can lead to fruitful insights into local conformational preferences within tau's unfolded ensemble.

IDPs as Targets for Drug Design

Due to increased life expectancies, the prevalence of age-associated dementias such as AD and DLB is projected to increase substantially [3]. Currently, available therapies for these diseases only slow the progression of disease or are palliative at best [125]. To date the FDA has approved five therapies - four cholinesterase inhibitors and an N-methyl D-aspartate (NMDA)

receptor antagonist – for AD [125]. Clearly there is a strong need for the development of novel therapies for these devastating diseases.

There are substantial data to suggest that aggregated forms of IDPs act as toxic species [9, 26]. Cell and animal models offer a tantalizing glimpse into the therapeutic value of disrupting aggregation promoting conformations or enhancing protein clearance [6, 7]. *In vitro* studies have identified peptides or small molecules which disrupt amyloid-β and tau aggregation, and that, in some cases, reduce cell toxicity [34, 35, 126, 127]. There are currently multiple clinical trials of therapies which aim to disrupt the aggregation of Aβ or tau protein in AD. Two of the most advanced drugs in clinical trials which target tau and Aβ, respectively, are methylionium chloride (Rember^{TM}), a tau aggregation inhibitor which is currently in phase II clinical trials and bapineuzumab, a passive immunization anti-Aβ monoclonal antibody in phase III clinical trials.

Like the study of IDPs themselves, the rational design of drugs targeting IDPs is a relatively new field. One of the key tools in rational drug design is the use of protein structure in the design strategy. The field of structure-based drug design (SBDD) aspires to design drugs based on the 3D structure of a target molecule [128]. Clearly this approach is inappropriate for a conformationally heterogeneous protein. Recently, however, Cheng et al proposed a systematic approach to targeting of IDPs for drug design [129]. Among other approaches, they suggest synthesizing a peptide mimic to a hydrophobic target region of the protein. This peptide can bind the IDP and may stabilize it in a unique conformation that may be suitable for crystallization. In this way, one can potentially isolate and study specific conformations that design molecules that help to prevent their aggregation. In principle, once aggregation-prone conformers are identified, existing structure-based design methods can be applied to design molecules that prevent the self-association of these problematic conformations [130, 131].

Chapter 3: Conformational Sampling with Implicit Solvent Models: Application to the PHF6 Peptide in Tau Protein

(This work was published as A. Huang and C. M. Stultz, "Conformational sampling with implicit solvent models: Application to the PHF6 peptide in tau protein," Biophysical Journal, vol. 92, pp. 34-45, Jan 2007.)

Abstract

Implicit solvent models approximate the effects of solvent through a potential of mean force and therefore make solvated simulations computationally efficient. Yet despite their computational efficiency, the inherent approximations made by implicit solvent models can sometimes lead to inaccurate results. To test the accuracy of a number of popular implicit solvent models, we determined whether implicit solvent simulations can reproduce the set of potential energy minima obtained from explicit solvent simulations. For these studies, we focus on a 6-residue amino-acid sequence, referred to as the paired helical filament 6 (PHF6), which may play an important role in the formation of intracellular aggregates in patients with Alzheimer's disease. Several implicit solvent models form the basis of this work – two based on the Generalized Born formalism, and one based on a Gaussian solvent-exclusion model. All three implicit solvent models generate minima that are in good agreement with minima obtained from simulations with explicit solvent. Moreover, free energy profiles generated with each implicit solvent model agree with free energy profiles obtained with explicit solvent. For the Gaussian solvent-exclusion model, we demonstrate that a straightforward ranking of the relative stability

of each minimum suggests that the most stable structure is extended, a result in excellent agreement with the free energy profiles. Overall, our data demonstrate that for some peptides like PHF6, implicit solvent can accurately reproduce the set of local energy minimum arising from quenched dynamics simulations with explicit solvent. More importantly, all solvent models predict that PHF6 forms extended β-structures in solution – a finding consistent with the notion that PHF6 initiates neurofibrillary tangle formation in patients with Alzheimer's disease.

Introduction

An appropriate representation of solvent is critical for obtaining physiologically relevant results from biomolecular simulations [132-134]. The most straightforward approach for modeling solvent is to explicitly include solvent molecules in molecular dynamics (MD) simulations. However molecular simulations with explicit solvent increase the degrees of freedom in the system and therefore can incur a significant computational cost. Consequently a number of implicit solvent models have been developed to reduce the computational complexity associated with solvated simulations. Such models modify the potential energy function to reproduce the effects of solvation without explicitly representing solvent atoms [132, 133]. As simulations with *implicit* solvent models have lead to important insights, these models have gained widespread acceptance in the field of biomolecular simulations [132]. As evidence of this, the literature is replete with studies that make conclusions based solely on data obtained from such models [132]. Recent studies, however, suggest that implicit solvent models can sometimes lead to results that are at odds with data obtained from explicit solvent simulations and experimental observations [135-137]. Therefore it is likely that not all implicit solvent

models are appropriate for every application. Moreover, the correct choice of solvent model to use for any given problem likely depends on the system to be studied, whether qualitative or quantitative results are desired, and the degree of accuracy required.

In the present study we explore whether conformational sampling with implicit solvent models can yield results similar to that obtained with explicit solvent simulations. The solvent models that form the basis of this work include: i) an early implementation of the Generalized Born (GB) model as described by Brooks and co-workers [138]; ii) an alternate implementation of the Generalized Born formalism which is based on an integral equation approach and that employs a simple smooth switching function (GBSW) [139]; iii) the effective energy function-1 (EEF1) implicit solvent model [140]; and iv) the TIP3P model of explicit solvent [141].

The GB model uses a linearized form of Still's equation to estimate the electrostatic component of the solvation free energy [138, 142]. The equation itself contains six independent parameters that are varied to optimize agreement between GB solvation energies and solvation energies calculated with a Finite-Difference-Poisson-Boltzmann (FDPB) algorithm [138]. As the Born radius is inversely related to the atomic polarization energy, Born radii can be calculated from the GB energies after parameter fitting [138]. The model has been widely applied and its utility has been demonstrated in a number of applications [138, 143].

The GBSW model, like the GB model, is based on Still's equation, however, GBSW employs a more rigorous integral equation approach to calculate the Born radii. In this method, the electrostatic solvation energy of a given atom is expressed as a sum of two terms – the self-solvation energy in the Coulombic approximation plus a term which accounts for the reaction field [139]. Each term is calculated using a surface/volume integration that employs a smooth switching function at the dielectric boundary to ensure numerical stability during molecular

simulations [139]. Unlike the GB method, the GBSW model contains two adjustable parameters which dictate the relative importance of the Coulombic field term and the reaction field term [139]. As before, the values of these parameters were obtained by minimizing the least square error between GBSW energies and those calculated with a FDPB approach [139]. Once the optimal values of the adjustable parameters are known, the Born radii can be calculated in a straightforward manner. The current implementation of the GBSW algorithm also incorporates a nonpolar contribution to the solvation free energy using the solvent exposed surface area of the protein of interest, and a user defined surface tension coefficient. The GBSW model has been used to refine model structures of the C-terminal domain of Hsp33 protein, obtained from sparse NMR data, into native-like folds which matched solved structures [144]. In addition, GBSW has been used to examine intermolecular interactions between actin and myosin, leading to new observations regarding a mutation associated with familial hypertrophic cardiomyopathy [145]. Overall, the model appears to be applicable to a broad range of problems.

The effective energy function-1 (EEF1) estimates the solvation free energy using a Gaussian solvent-exclusion model [140]. EEF1 expresses the solvation free energy of a protein as a sum of group contributions, where each contribution is equal to a reference solvation energy (i.e., the solvation energy of the group alone) minus an integral over a solvation free energy density function. The underlying assumption is that the integral over the free energy density is well approximated by a sum of Gaussian functions [140]. Important aspect of the model are that charged side chains are neutralized and a distance dependent dielectric is used to further attenuate electrostatic interactions. The model has been used in a number of applications and interesting results have been obtained. Most notably, EEF1 has been used to calculate unfolding trajectories of proteins [146], discriminate correctly folded from unfolded structures [147], and to

probe the interactions between regions of α-lytic protease (αLP) leading to a better understanding of the relative importance of different interactions in stabilizing the native state [148].

In the present study we address a specific, well-defined, problem. We determine whether each of these solvent models can reproduce the set of local energy minima obtained from quenched MD (QMD) simulations with explicit solvent. To this end we perform QMD simulations with each of the aforementioned implicit solvent models and compare these results to those obtained with a TIP3P model of solvent. We note that QMD is a widely used method for locating local energy minima on a given potential surface. The procedure consists of high temperature MD simulations (typically at 1000K), followed by minimization of the resulting structures [149]. High temperature simulations ensure that a wide region of conformational space is sampled and the subsequent minimizations assure that only local energy minima are analyzed. Minimization can be performed by coupling the system to a heat bath at 0K [150, 151], or by using standard energy minimization algorithms such as steepest descent or conjugate gradients [152]. QMD has been used to determine optimal positions and orientations of small functional groups in the binding site of an enzyme [150], estimate the density of states for proteins [153], and to study the conformational landscape of peptides and peptide analogues [151, 152].

Our studies focus on a 6-residue peptide commonly referred to as paired-helical filament 6 (PHF6), which corresponds to the sequence found at the N-terminus of the third microtubule-binding repeat domain of tau protein (^{306}VQIVYK^{311}). Tau protein forms intracellular aggregates (also known as neurofibrillary tangles) in patients with Alzheimer's disease (AD) and PHF6 corresponds to the minimal region of tau needed for aggregation to occur *in vitro* [5, 154, 155]. As the formation of intracellular aggregates may be responsible, in part, for neuronal death in

patients with AD, the predominant low energy states of PHF6 are of particular interest [114, 156]. In performing an analysis of PHF6, the goals of this work are not only to evaluate the ability of several implicit solvent models to reproduce energy minima on a potential surface that explicitly models solvent, but also to determine the most stable conformations of this peptide.

Methods

Quenched Molecular Dynamics with Explicit Solvent

Quenched molecular dynamics consisted of high temperature MD followed by extensive minimization of the structures sampled during the trajectory. A polar hydrogen model of the PHF6 peptide (VQIVYK) was created from the CHARMM19 polar-hydrogen parameter set and initial coordinates for PHF6 were built using the IC facility, all within CHARMM [157]. Both the N and C-termini of the peptide were patched using NTERM and CTERM patches, as is commonly done, resulting in charged termini (i.e., $-NH3^+$ and $-COO^-$). The resulting structure was solvated with an equilibrated set of TIP3P water molecules, and waters that overlapped with the peptide or that were outside of a 19Å radius were removed. A total of 823 water molecules were added to the system. A stochastic boundary setup with a solvent sphere of radius 19Å was used for these simulations [158]. The system was minimized, then heated and equilibrated for 1ns at 1000 K. Production dynamics were performed for an additional 10ns at 1000 K. Sampling at this temperature facilitates a broad exploration of the conformational space. The temperature was maintained by weakly coupling (tcoup = 5ps) the system to a heat bath using the Berendsen method [159]. All explicit solvent simulations employed a nonbond interaction cutoff of 17Å electrostatic interactions were shifted to zero between 14Å to 16Å while a switching function

was used to cutoff van der Waals interactions at 16Å. SHAKE was used to hold hydrogen bond distances close to their equilibrium values and a 2fs time step was used [160].

Structures were chosen from the trajectory every 10ps and subsequently minimized, resulting in 1000 distinct minimum energy structures. Minimizations were performed on the entire system consisting of the peptide and all explicit water molecules. In addition, minimizations used the nonbond specifications outlined above and consisted of 2500 steps of steepest descent followed by 2500 steps of conjugate gradient minimization. A root-mean-square gradient (GRMS) cutoff of 0.01kcal/mol/Å was set, such that if the system achieved a GRMS below this value during the minimization protocol, then the minimization was terminated. The procedure for heating, equilibration, sampling, and minimization was identical for all of the solvent models investigated in this study.

Quenched Molecular Dynamics in vacuum

Quenched molecular dynamics simulations were performed in vacuum (ε=1). Comparing the vacuum minima with minima obtained with the different solvent models enabled us to assess the affects of the solvent models on the structure of the peptide. The nonbond cutoffs and the minimization protocol were identical to those used in the explicit solvent simulations.

Quenched Molecular Dynamics Simulations with Implicit Solvent

We performed a similar procedure for finding local energy minima on the potential energy surface of each implicit solvent model described above. One issue that needs to be resolved is the correct choice of simulation conditions for each implicit solvent model. In general, we rely on prior data to choose simulation conditions that optimize the chance that each implicit solvent

simulation would reproduce minima obtained from the explicit solvent simulations. In this regard, we note that some temperature coupling algorithms may not be appropriate for all implicit solvent simulations [161]. In explicit solvent simulations with a Berendsen heat bath, the entire system, consisting of both the solute and the solvent, are coupled to an external heat bath. Implicit solvent simulations that utilize similar thermostats only couple the peptide to an external bath as a continuum model is used for solvent. It has been noted that some thermostats which couple the solute alone to a heat bath may lead to diminished atomic fluctuations, especially when the peptide itself is tightly coupled [161]. Diminished rms fluctuations would clearly be disadvantageous for an approach which attempts to map local energy minima on a large potential surface.

In order to determine whether a Berendsen thermostat with a coupling constant of 5ps would be appropriate for our studies, we conducted MD simulations of PHF6 with each implicit solvent model outlined above and compared these data to simulations conducted with explicit solvent (when both the peptide and solvent are coupled to an external bath). The resulting root-mean-square (rms) fluctuations were then compared to rms fluctuations arising from the explicit solvent simulations. For PHF6 the rms fluctuations arising from all of the implicit solvent simulations are in reasonable agreement with the rms fluctuations from the explicit solvent simulations (Figure 5). As we are primarily interested in mapping the local energy minima on the different potential energy surfaces, and not the dynamical properties of PHF6 in different models of solvent, these data suggest that simulations with a Berendsen thermostat would be appropriate for our studies.

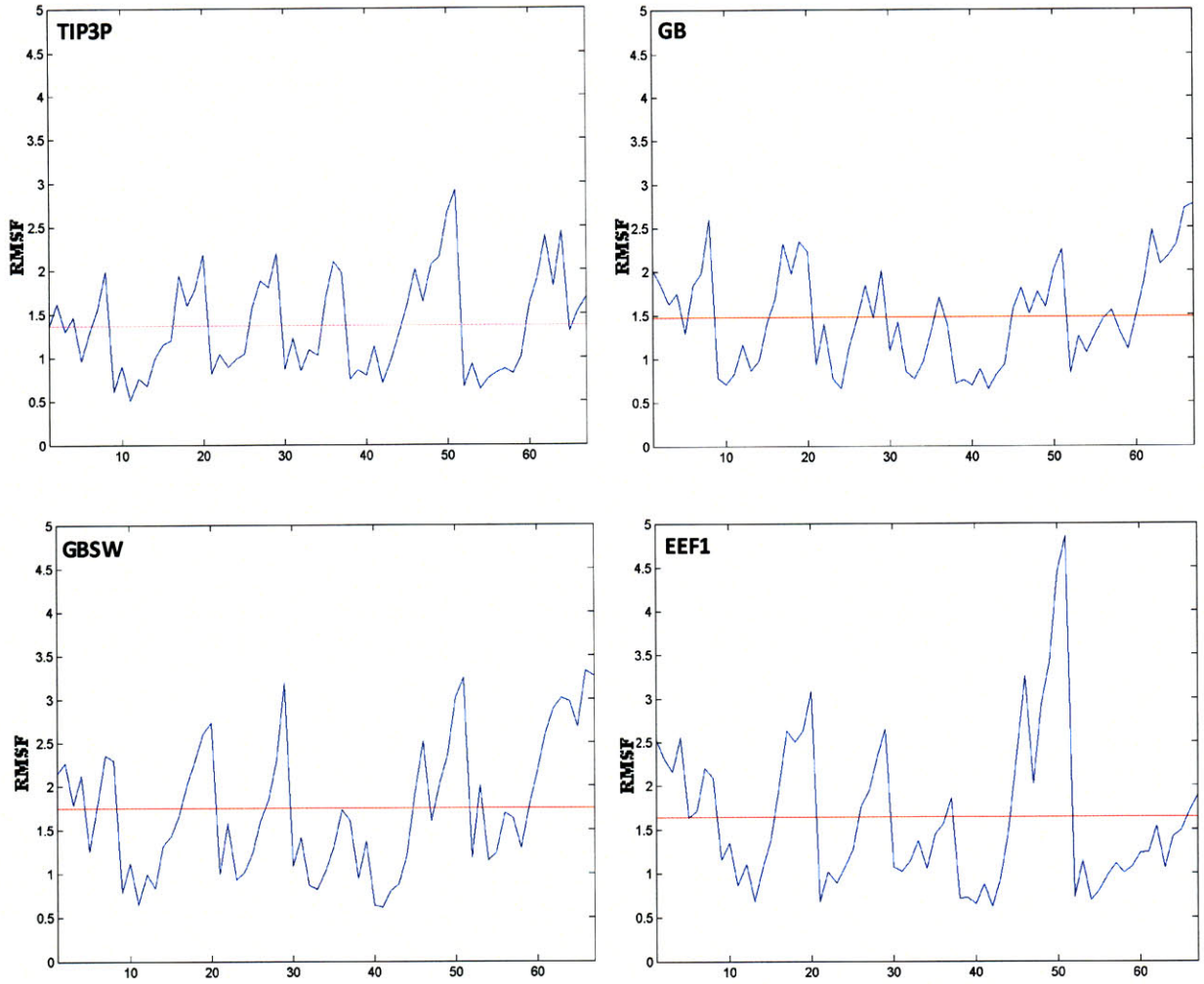


Figure 5 RMS fluctuations for 100ps simulation (100ps equilibration, 100ps production dynamics all at 300K) of PHF6 in different solvent models using a Berendsen heat bath. Simulation parameters, including nonbond cutoffs, are as listed in Methods. The average rms fluctuation over all atoms is denoted with a red line.

Lastly we note that the precise model for the nonbond interactions for each implicit solvent model was chosen based on prior data. The goal here was to optimize the chance that each model would produce data in agreement with the explicit solvent results.

Quenched Molecular Dynamics with GB

Generalized Born simulations utilized the implementation, and Born radii, originally described by Dominy et al. [138]. As in our previous study [137], no truncation of nonbond

terms was used as this approach yields better results relative to an approach that employs finite nonbond cutoffs [162].

Quenched Molecular Dynamics with GBSW

GBSW simulations used the implementation previously described by Im et al. with a half smoothing length of 0.3Å, a nonpolar surface tension coefficient of 0.03 kcal/(mol x $Å^2$), and a grid spacing of 1.5Å [139]. Nonbond cutoffs were set to 16Å using a switching function for both van der Waals and electrostatic interactions. Of the 1000 structures, the minimization protocol described above failed for a single structure, which was excluded from the analysis, yielding 999 distinct structures. The singular failed structure was in a non-physical conformation corresponding to an energy of 1.3×10^{11} kcal/mol, while all other structures had energies that were less than -400 kcal/mol.

Quenched Molecular Dynamics with EEF-1

The EEF-1 implicit solvent model was used as implemented in CHARMM [140, 157]. As the nonbond cutoff parameters are integral to the model, the previously described nonbond cutoffs were used here.

Generation of Ramachandran Plots

Ramachandran density surfaces were created from the minima generated from each of the quenched dynamics simulations. The φ/ψ values for residues Gln2-Tyr5 were calculated for each of the 1000 minima (999 for GBSW), and a density function was computed using the scattercloud function (written by Steve Simon) obtained from the MATLAB central code repository (*http://www.mathworks.com/matlabcentral/*). The densities were normalized by their

maximum values and rendered as surface plots using MATLAB. Approximate secondary structure regions as defined in [1] corresponding to α-helical and β structure are colored.

Generation of Minimum Pairwise Distance (MPD) Plots

Histograms of Minimum Pairwise backbone rms Deviations (MPD) between minima from different models (a reference model and a comparison model) were computed. These histograms were used to determine whether each minimum in the reference model was adequately represented by a structurally similar minimum in the comparison model. For example, suppose explicit solvent is the reference model and data arising from the EEF1 simulations is the comparison model. The MPD plot is used to determine if each explicit solvent minimum is represented in the set of EEF1 minima. For each TIP3P minimum, we find the EEF1 minimum with a backbone conformation closest to the TIP3P minimum in question. This set of rms deviations provides an objective assessment of how well the EEF1 minima reproduce the structures corresponding to the explicit solvent minima. It is also of interest to determine the converse; i.e., whether each EEF1 minimum is well represented by an explicit solvent minimum. The converse is computed by setting EEF1 as the reference model and the explicit solvent results as the comparison model. If EEF1 generated many spurious minima that did not correspond to explicit solvent minima, then the resulting histogram of rms deviations would contain many large values. Therefore, two sets of MPD plots were computed for each of the implicit solvent models. In one set of calculations the explicit solvent minima formed the reference set and in the other set of calculations, the implicit solvent model served as the reference. Histograms were computed using MATLAB and plots of aligned structures were constructed with VMD [163].

Potential of Mean Force Calculations for PHF6

Free-energy profiles for PHF6 were computed for each solvent model. The reaction coordinate for these simulations was the radius of gyration of the peptide main-chain atoms. The simulations began by restraining the backbone to adopt an extended conformation with a radius of gyration of 5.5Å using a harmonic constraining potential with a force constant of 25 kcal/mol/Å^2. The system was then equilibrated at 300°K for 1ns. The potential of mean force (pmf) for a given solvent model was calculated by running a series of simulations (windows), where the peptide is restrained to a different radius of gyration using a harmonic force constant of 25 kcal/mol/Å^2. The first window was centered at 5.5Å and subsequent windows began with the final state from the preceding window. The radius of gyration was decreased by 0.1Å for each new window. Restrained molecular dynamics for each window involved 20ps of equilibration followed by 80ps of production dynamics. Additional dynamics were performed to extend the pmf boundaries and improve sampling for regions of the pmf that exhibited discontinuities. Specifically, windows for extended states of the peptide were run at 0.1Å intervals for rgyr constraints ranging between 5.6Å and 6.6Å to extend the boundaries of the pmf.

In order to compute the potential of mean force, the radius of gyration was computed every 20fs for each window of dynamics. From these data a biased probability density, ρ_i^* is computed and the potential of mean force, $W_i(\xi)$, is computed using the relation [158]:

$$W_i(\xi) = -k_B T \ln \rho_i^*(\xi) - V_i(\xi) + C_i \qquad (3.1)$$

where k_B is Boltzmann's constant, T is the temperature, V_i is the restraining potential for window i, and C_i is a constant. In order to construct one continuous potential of mean force, the

different pmfs from each window need to linked together – a process performed by the program SPLICE [164].

In order to determine that our pmf had converged, we performed additional simulations for new windows constrained at RGYR that were offset 0.05Å from the original window constraints and determined that this convergence criterion was satisfied. Our metric for convergence of the pmf was based on the location of the pmf minimum, since this is the primary quantity of interest for this study. Specifically, we required that the location of the pmf minimum changed by less than 0.25Å as the window step size was halved.

Representative structures from the global energy minimum in each pmf were generated by first averaging the structures sampled at the window corresponding to the global energy minimum followed by minimization to the nearest local energy minimum. All molecular figures were constructed with VMD [163].

Calculating Vibrational Entropies

Vibrational entropies were calculated from the 1000 distinct minima obtained from EEF1 simulations. To ensure that only non-negative eigenvalues would be generated from the normal mode calculations, each minimum was further minimized using 1000 steps of steepest descent minimization followed by 2500 steps of adopted-basis Newton Rhapson minimization. The corresponding Hessian matrix was then diagonalized to yield the normal modes and their corresponding frequencies. The vibrational entropy for a given minimum was computed as follows [165, 166]:

$$-TS_{vib} = \sum_{i=1}^{3N-6} \left(k_B T \ln(1 - e^{-h\nu_i/(k_B T)}) - \frac{h\nu_i}{e^{h\nu_i/(k_B T)} - 1} \right) \qquad (3.2)$$

where N is the number of atoms in the system, h is Planck's constant, k_B is Boltzmann's constant, and $\{\nu_i\}_{i=1}^{3N-6}$ are the normal mode frequencies. CHARMM, was used to create the Hessian matrix from minimized structures and MATLAB (© Mathworks) was used to calculate vibrational entropies from the Hessian matrix, yielding 1000 vibrational entropy measures; i.e., one for each minimum [157].

We note that a harmonic analysis could only be performed on minima arising from the EEF1 simulations as second derivative calculations with GB are not supported in CHARMMv32a2 and despite the extensive additional minimization, Hessian matrices for GBSW structures had negative eigenvalues, thereby preventing a normal mode analysis.

Results

Minimum energy conformations with explicit solvent

Minima on the potential energy surface of PHF6 were obtained from high temperature molecular dynamics simulations with explicit solvent followed by extensive minimization (i.e., quenched dynamics). After 10ns of molecular dynamics at 1000K, a range of conformations was sampled and subsequent energy minimization yielded 1000 distinct structures corresponding to different local energy minima. These structures span a range of conformations from the compact, with a radius of gyration (rgyr) near 3Å, to a rgyr of almost 5.6 Å (Table 1). By contrast, minima arising from quenched molecular simulations in vacuum are relatively homogeneous and have radii of gyration that are distributed over a narrow range – between 3.0Å and 3.5Å, suggesting that compact states are overwhelmingly favored in the vacuum simulations (Table 1).

Solvent Model	Average rgyr (Å) ± std	Minimum rgyr (Å)	Maximum rgyr (Å)
TIP3P	4.1 ± 0.63	3.0	5.6
Vacuum	3.1 ± 0.06	3.0	3.5
GB	4.7 ± 0.46	3.2	5.7
GBSW	4.5 ± 0.55	3.1	5.7
EEF1	4.7 ± 0.47	3.4	5.9

Table 1: Statistics of minima obtained from quenched molecular dynamics simulations with different solvent models.

To quantify the diversity among the different minimum energy structures, we computed the backbone rms deviation between all pairs of minima (Figure **6**). As we are interested in distinguishing extended structures from compact structures, in addition to secondary structural motifs sampled by the peptide, we focus on comparisons of the backbone rms deviation between different pairs of conformers. These data confirm that the explicit solvent minima are considerably more diverse than minima arising from the vacuum simulations. In particular, the most extended structure from the vacuum simulations has a radius of gyration of only 3.5 Å and contains a salt bridge between the N and C-termini (Figure 7). In vacuum this salt bridge is exceptionally stable in that it has an interaction energy near -90kcal/mol and remains intact even at 1000K. Hence virtually all minima have this salt-bridge and the resulting vacuum structures are all compact.

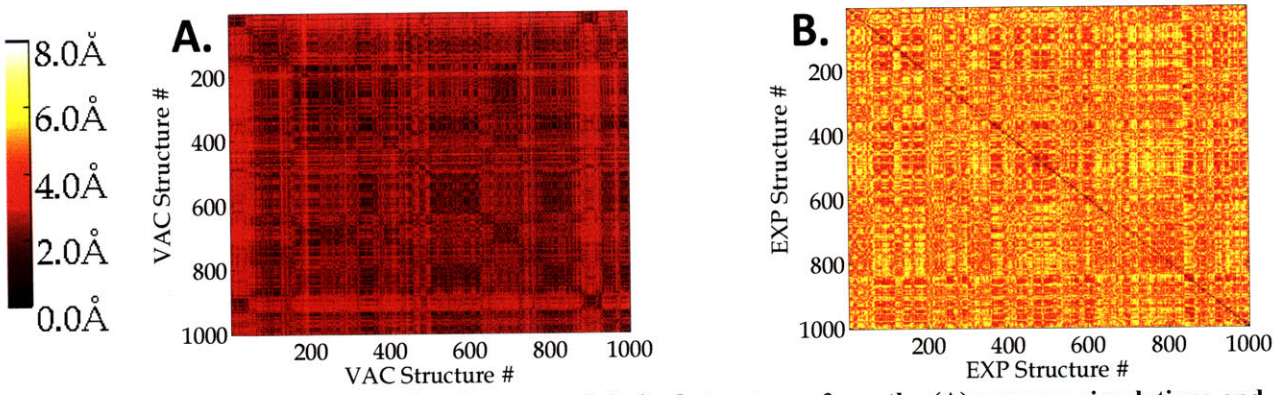


Figure 6: Pairwise distance matrices between minimized structures from the (A) vacuum simulations and (B) explicit solvent simulations. Each pixel color corresponds to a pairwise backbone RMS distance. The color scale is shown at the left of the figure.

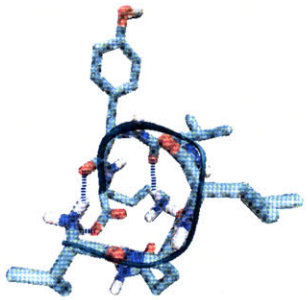


Figure 7: Structure of the most extended PHF6 minimum arising from the vacuum simulations (rgyr=3.5Å).

Minimum energy conformations with implicit solvent

Minima arising from QMD simulations with implicit solvent sample a range of radii of gyration that is similar to that found in the set of explicit solvent minima (Table 1). A comparison between representative minima from the different solvent models further illustrates the close correspondence between the implicit solvent results and the explicit solvent results (Figure 8); i.e., the backbone conformations of the implicit solvent minima are similar to that arising from the explicit solvent simulations.

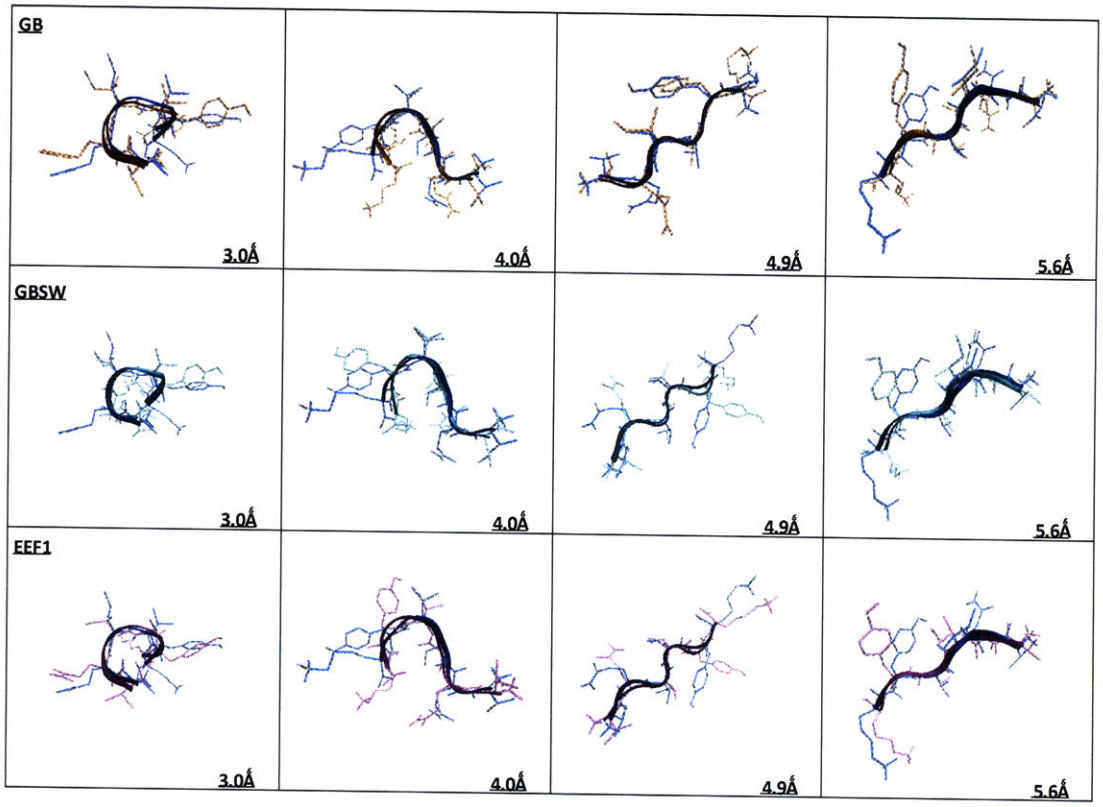


Figure 8: Representative explicit solvent structures (blue) aligned with their closest implicit solvent structures. The first row depicts the alignment of GB (orche) minima to TIP3P minima; the second row shows the alignment of GBSW (cyan) minima to TIP3P; and the last row shows the alignment of EEF1 (purple) minima to TIP3P minima.

The degree of similarity between the implicit solvent minima and the TIP3P minima was quantified by computing minimum pairwise distance (MPD) plots. Each MPD plot is a histogram of the minimum pairwise backbone rms deviations between minima from two different models; a reference model and a comparison model. For each minimum in the reference model, the closest minimum in the comparison model is found and used to generate a histogram of rms deviations. For example, in Figure 9A the TIP3P minima is the reference model and the GB minima is the comparison model. These data demonstrate that every explicit solvent minimum is within 1.5Å of a GB minimum (Figure 9A).

A also shows an overlay of the explicit solvent minimum that is farthest away from a GB minimum; even for this worst case, the two minima have very similar backbone conformations.

MPD plots for the other implicit solvent models reveal the same trend; i.e., each explicit solvent minimum is within 1.3Å of a GBSW minimum (Figure 9B) and 1.5Å from an EEF1 minimum (Figure 9C).

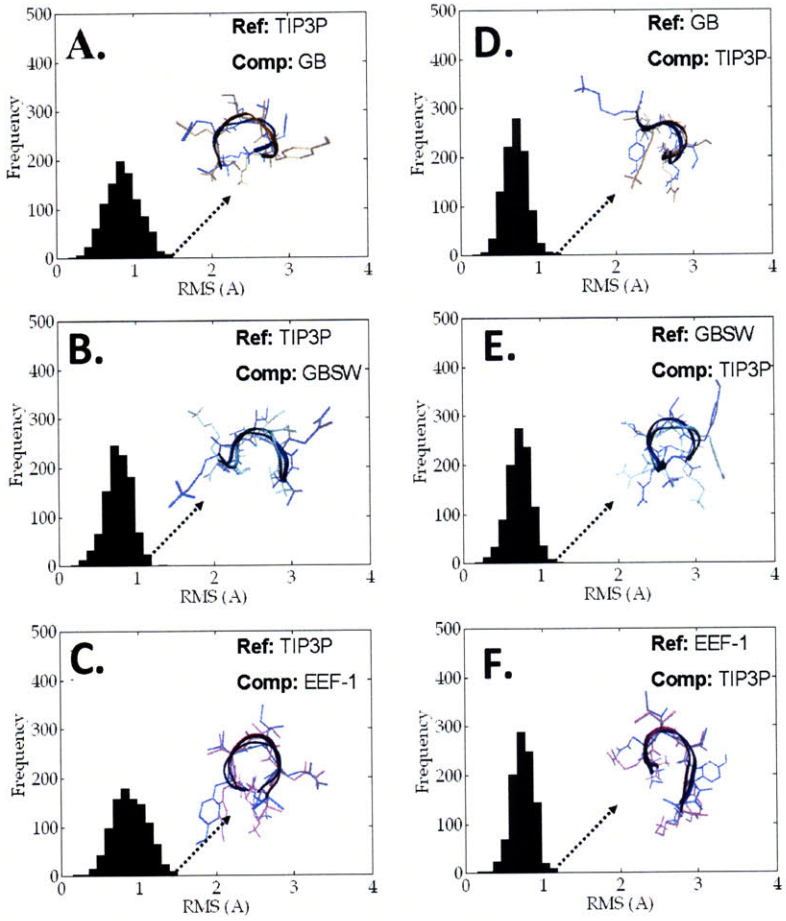


Figure 9: Minimum Pairwise Distance (MPD) plots (see text). The reference and comparison sets are labeled. In each case, the two structures having the greatest RMS difference is overlaid.

While every explicit solvent minimum is close to an implicit solvent minimum, it may be that the implicit solvent simulations produce extraneous minima that do not correspond to any minimum arising from the TIP3P simulations. To determine whether the implicit solvent simulations produced such superfluous minima, the reverse comparison was done; i.e., MPDs were computed with each implicit solvent minima serving as the reference model and TIP3P serving as the comparison model (D-F). These data verify that the implicit solvent simulations do not produce many extraneous minima – that is, each implicit solvent minimum is close to an explicit solvent minimum.

A conformational analysis of the TIP3P minima suggests that the four residues in PHF6 with defined φ/ψ angles (residues 2-5) preferentially sample regions of conformational space corresponding to β-structure (Figure 10). Gln2, in particular, is most likely to adopt φ/ψ angles belonging to the β strand region of conformational space. The φ/ψ densities of the GB, GBSW and EEF1 minima are similar to that obtained from the TIP3P simulations in that β-strand configurations are also favored (Figure 10). By contrast, the vacuum simulations yield minima where residues 2, 4 and 5 adopt φ/ψ angles that belong to the α-helical region of conformational space (Figure 10).

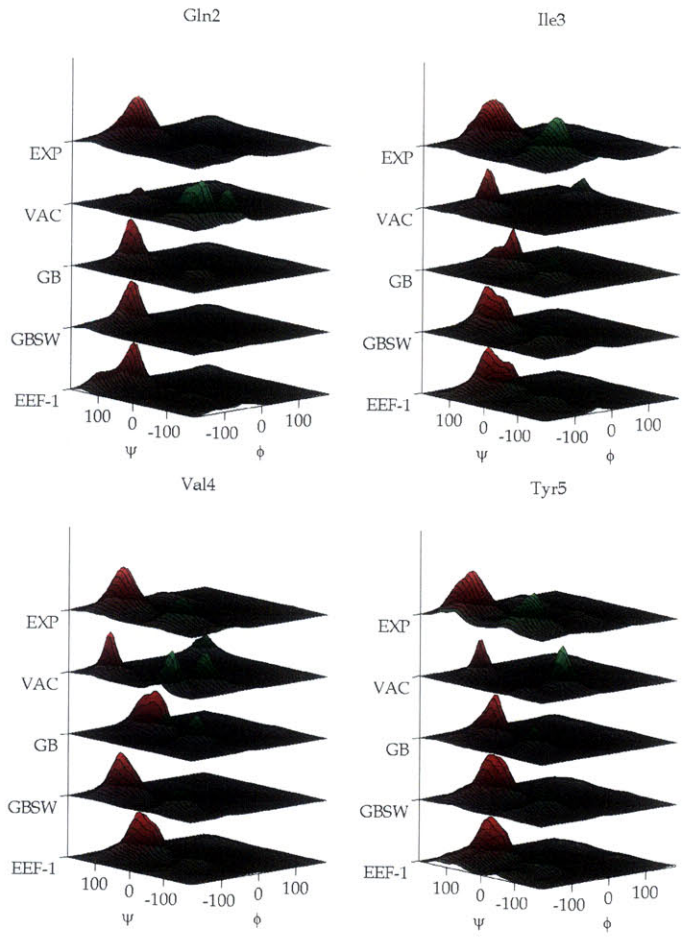


Figure 10: Comparison of normalized φ/ψ densities of minima obtained by quenched molecular dynamics for residues Gln2-Tyr5. The region corresponding to the β-structure peak is colored red and the region corresponding to the alpha-helix peak is colored green. Following the general secondary structural regions used in [1], the region of β-sheet conformations consists of φ/ψ angles within the range of φ=[-180°,-45°] and ψ=[45°,225°] and the region of α-helix conformations consists of φ/ψ angles within the range of φ=[-180°,0°] and ψ=[-100°,45°].

Potential of Mean Force Calculations

Free energy profiles were calculated for PHF6 in explicit solvent to determine the predominant conformation of the peptide in solution (Figure 11). The reaction coordinate for these simulations was the radius of gyration of the peptide. The global free energy minimum of the peptide in explicit solvent occurs at approximately 5.2Å, corresponding to a relatively

extended conformation of the peptide (Figure 11) – a finding consistent with the φ/ψ densities of explicit solvent minima.

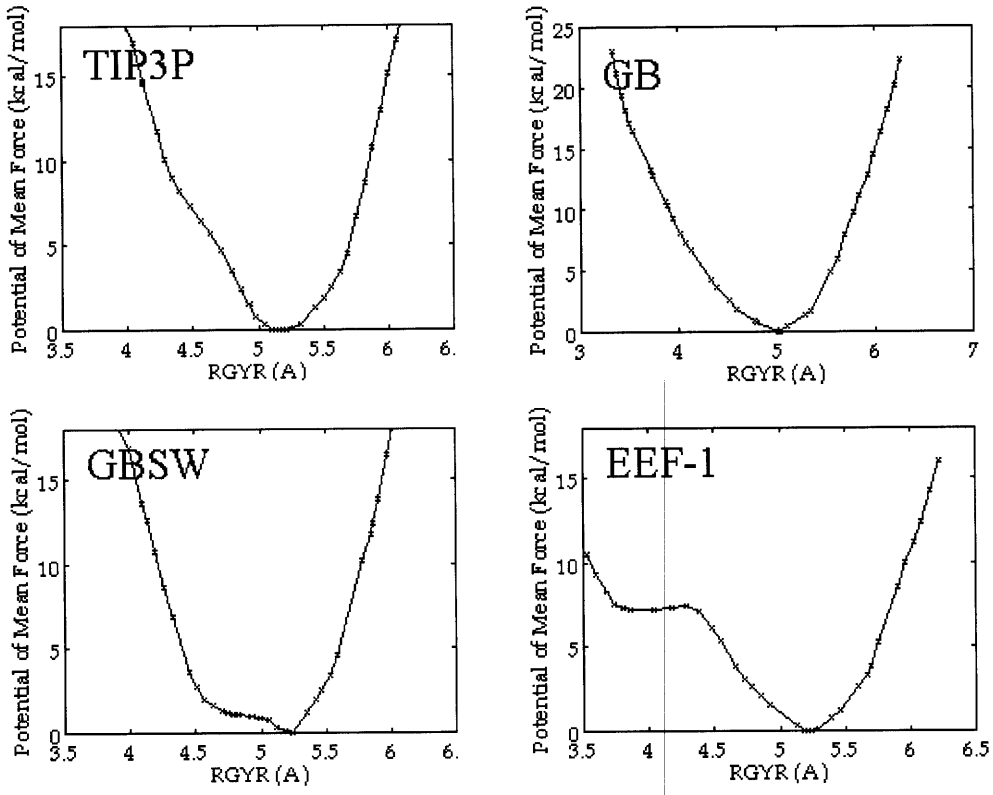


Figure 11 Potential of mean force plots for the different solvent models analyzed in this study.

The free energy profiles calculated with each of the implicit solvent models are similar to the pmfs calculated with explicit solvent; i.e., each has a global minimum located between 5Å and 5.5Å (Figure 11). Average structures from windows corresponding to the pmf minima confirm that these low energy structures are relatively extended (Figure 12). In addition, residues 2-5 from the average structure arising from the explicit solvent pmf minimum have φ/ψ angles that fall within a region of conformational space consistent with β-structure. The GBSW average structure, however, is least similar to the average structure from the explicit solvent pmf minimum (Figure 12). The backbone rms deviation between the GBSW pmf minimum and the TIP3P pmf minimum is approximately 2.7Å, whereas the GB and EEF1 structures are within 1Å

of the TIP3P pmf minimum structure (Figure 12). Hence while all of the implicit solvent models show qualitative agreement with the explicit solvent pmf, the average structure arising from EEF1 simulations at the global free energy minimum is most similar to the average structure obtained from corresponding simulations with explicit solvent.

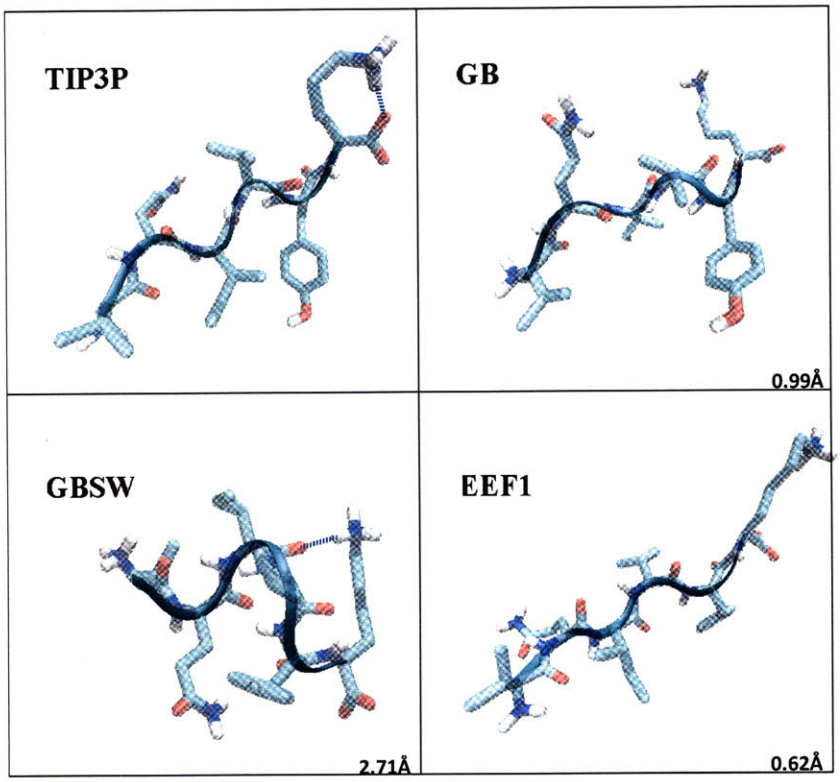


Figure 12: Representative structures from the simulation windows corresponding to the global free energy minimum in each pmf. The backbone rms deviation from the TIP3P structure is explicit shown for each of the implicit solvent structures.

Ranking Minima from the Implicit Solvent Models

Ideally any sampling protocol designed to find low energy states on a potential surface should not only discover local energy minimum, but it should also deduce which of the resulting low energy structures are the most stable. In this regard, we note that EEF1 and potentials based on the generalized born formalism have been shown to correctly identify the most stable protein conformation from sets consisting of native and misfolded structures [147, 167-169]. Moreover,

a number of these studies suggest that the most stable state can be deduced from static energy calculations on energy-minimized structures [147, 167, 169]. Given these observations, we explored whether static energy calculations on the different implicit solvent minima could provide enough information for identifying the most stable conformation.

A comparison of the relative energies of the different minima is shown in Figure 13. Both the GB and GBSW minima have a number of low energy states that are within 2kcal/mol of the lowest energy structure, and all of these conformations are relatively compact with radii of gyration near 3.5Å (Figure 13). By contrast, the set of EEF1 minima contains a prominent minimum with a radius of gyration of 5.08Å, a value close to the global free energy minimum in the EEF1 and TIP3P free energy profiles (Figure 13). Hence the most stable conformation of PHF6 can be identified from an analysis of the EEF1 energies alone.

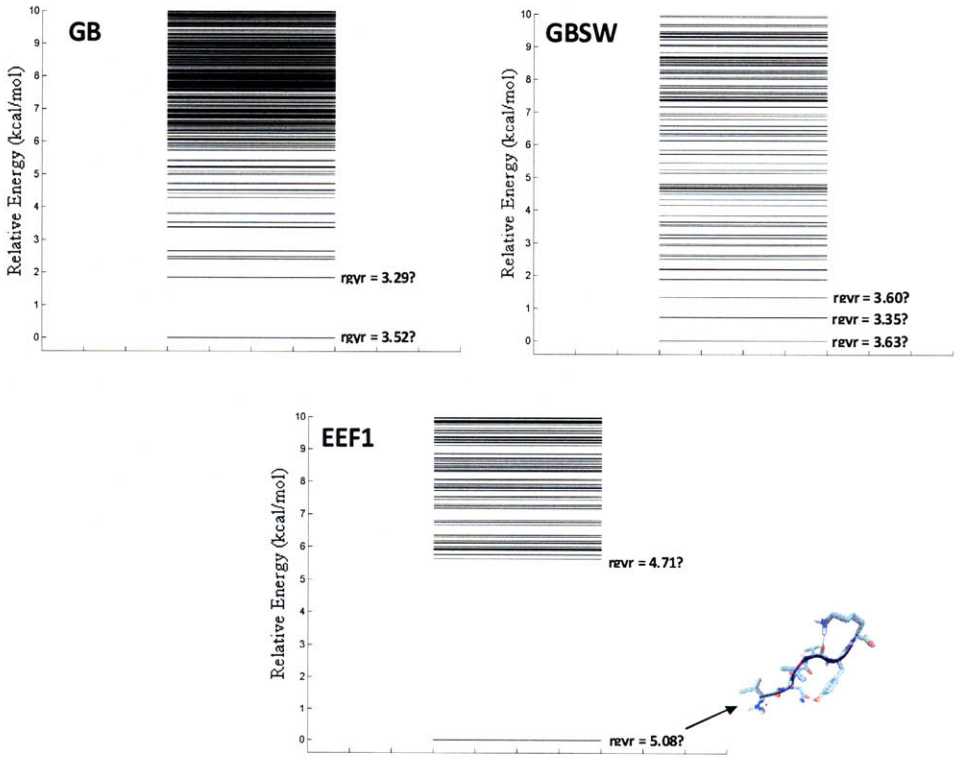


Figure 13: Relative energies of minima from each implicit solvent simulation. The radii of gyration of the low energy structures in each solvent model are explicitly shown. The structure of the lowest energy minimum arising from the EEF1 simulations is explicitly shown.

We note that methods which identify the most stable conformation of a protein from static energy evaluations on distinct energy-minimized conformers typically assume that the solute entropy at each local energy minimum is roughly the same, and therefore can be ignored [168, 170, 171]. Such approximations may be valid for a number of proteins, but it is not clear whether such a premise is valid for small peptides like PHF6 [170, 171]. While static energy calculations with EEF1 lead to results that agree with calculated free energy profiles, this does not necessarily imply that the solute entropy is the same at each minimum. Therefore, to explore the role that the solute entropy has in determining the relative stability of the PHF6 minima, we computed the vibrational entropy of each EEF1 minimum within the context of a harmonic approximation [166]. The relative free energy of each minimum was then estimated using the sum of the internal energy (i.e., the EEF1 energy) and the vibrational entropy (Table 2). Ranking the EEF1 minima using this new measure leads to conclusions that are identical to what was obtained from an analysis of the EEF1 energies alone. In particlar, the lowest energy conformations are extended, and the lowest energy structure is the same (Table 2). However, as is clear from Table 2, the vibrational entropy spans a range of more than 10kcal/mol, a somewhat larger range than was noted in prior studies on proteins [168, 169]. Including the vibrational entropy also leads to a change in the ranking of the PHF6 minma. Consequently, even though our results are similar to those seen when the vibrational entropy is explicitly included, it is clear that it can play a role in determining the relative ordering of different minima.

Ranking	rgyr (Å)	E (kcal/mol)	$-TS_{vib}$ (kcal/mol)	$A = E - TS_{vib}$ (kcal/mol)
1	5.08	-191.93	-23.22	-215.15
2	4.90	-177.75	-33.31	-211.06
3	5.16	-185.81	-24.92	-210.73
4	5.03	-186.17	-24.53	-210.70
5	5.10	-185.98	-23.82	-209.80
15	5.04	-186.03	-22.63	-208.66
38	5.01	-177.09	-30.60	-207.69
64	4.45	-185.74	-20.99	-206.73
141	4.46	-172.93	-32.45	-205.37
843	3.99	-175.72	-20.53	-196.25

Table 2: EEF1/Vibrational energies of selected EEF1 minima. Minima are ranked in order of increasing energy.

Discussion

Given their considerable computational efficiency, a number of problems can be approached with the aid of implicit solvent models that would be intractable if only explicit solvent models were available [143-148, 167]. However not all implicit solvent models are created equal, and some may be more appropriate for particular problems. As such, studies such as the present work, which aim to delineate the limitations as well as the advantages of different implicit solvent models, may help to decide which model to use for any given application.

This study was designed to address a specific question – namely, could selected implicit solvent models adequately reproduce the set of local energy minima found on a potential surface that explicitly includes solvent. Towards this end, we mapped local energy minima on different potential surfaces and compared these minima to minima obtained from simulations with explicit solvent. We found that GB, GBSW, and EEF1 performed quite admirably in that they were able to successfully reproduce the set of minima obtained from explicit solvent simulations. Ramachandran plots of the resulting structures confirm that all solvent models sampled similar regions of conformational space. Furthermore, free energy profiles obtained from all three implicit solvent models were in good agreement with free energy profiles obtained with explicit

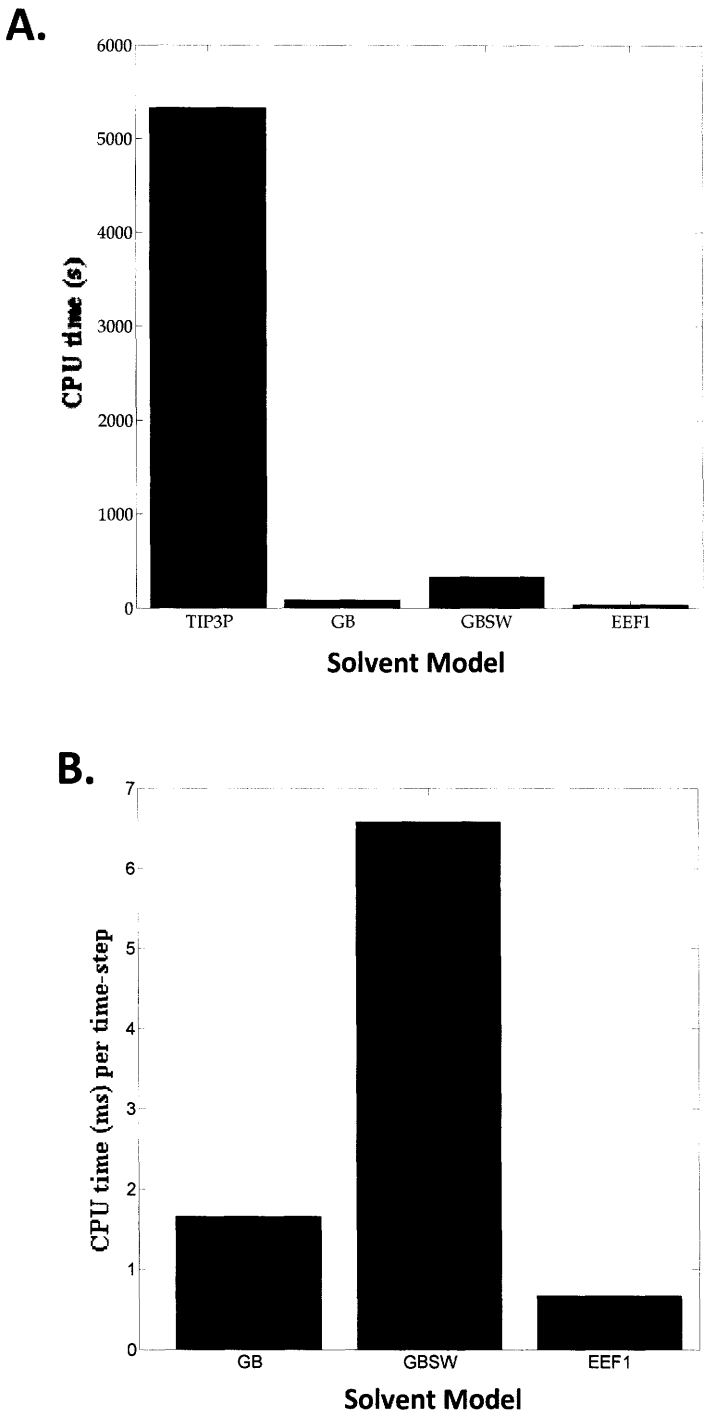


Figure 14: (A) CPU time for running one window of pmf simulations in each solvent model. (B) Close up of CPU requirements for the various implicit solvent models. All calculations were performed on one XEON 2.8GHz processor running Linux.

solvent. However, visual inspection of the structures suggests that EEF1 provides a slightly more accurate representation of the most favored conformations on the peptide's free energy surface.

All of the implicit solvent simulations generate pmfs that are in good agreement with the explicit solvent simulations in a fraction of the CPU time required for the explicit solvent simulations (Figure 14A). Of the different implicit solvent simulations, EEF1 required the least CPU time (Figure 14B). This is due, in part, to the different nonbond cutoffs in each model. As the nonbond specifications in EEF1 are part of the model, all EEF1 simulations employ a relatively short cutoff of 9Å [140]. The nonbond cutoffs for the GB and GBSW models were considerably larger. The GB simulations employed an infinite cutoff because it has been shown that this cutoff scheme yields results that are in good agreement with explicit solvent for some systems [162]. The GBSW

simulations used a finite nonbond cutoff of 16Å because this value leads to reasonable computation times with relatively small errors in the calculated forces [139]. Nevertheless, a 16Å cutoff for a small peptide like PHF6, leads to almost no truncation of the nonbond terms. As a result, the nonbond lists for the GB and GBSW simulations are quite similar. The longer simulation time for GBSW is due to the fact that, unlike GB, GBSW employs a relatively expensive surface/volume integration to calculate the electrostatic contribution to the solvation energy [138, 139].

In order to determine whether the most stable state of PHF6 could be identified from an analysis of the minima alone without additional umbrella sampling, we examined the relative energies of minima arising from each implicit solvent simulation. The lowest energy structure from the set of EEF1 minima is extended and has a radius of gyration near that found in the free energy profiles. By contrast, the lowest energy structures from the GB and GBSW simulations are relatively compact. Hence, for PHF6, one could correctly deduce that extended structures are most stable from an analysis of the EEF1 energies alone. These data are encouraging as they suggest that an analysis of minima obtained from simulations with EEF1 may provide insights that are comparable to what one would obtain from umbrella sampling calculations with explicit solvent – a considerably more taxing approach.

It should be noted that this conclusion may not be generally applicable. Ranking EEF1 minima based solely on static EEF1 energies assumes that the solute entropy at each minimum can be safely ignored. However, estimates of the vibrational entropy reveal that the solute entropy can vary significantly at each minimum. Although our conclusions are the same when the vibrational entropy of each minimum is explicitly calculated, the ranking of the EEF1 minima is somewhat altered when this is done. Therefore we cannot rule out that estimates of

the solute entropy are needed to accurately identify the most stable conformation of other peptides. In this regard, we note that static energy evaluations of GB and GBSW minima lead to conclusions that differ from that obtained from the pmf calculations in explicit solvent. As normal mode analyses could not be performed on GB and GBSW minima, it may be that more accurate results could be obtained if a vibrational analysis were performed on these minima.

In our previous study we found that both EEF1 and GB were unable to reproduce the free energy profile obtained from simulations with explicit solvent using a different peptide system [137]. In that work we umbrella sampling calculations with explicit solvent to calculate this peptide's potential of mean force as a function of its radius of gyration [137]. The FRET efficiency for this peptide, which was calculated from the pmf, was in excellent agreement with experiment. Central to the success of the explicit solvent simulations was the formation of a stable salt bridge between glutamate 5 and arginine 11. By contrast, in both the GB and EEF1 simulations, the formation of a glutamate-arginine salt bridge was unfavorable, and consequently simulations with these implicit solvent models lead to calculated FRET efficiencies that disagreed with the explicit solvent results [137]. While the solvation energy of individual side chains is likely well modeled by these implicit solvation models, it is not clear that energetics of salt-bridge formation is appropriately modeled by these approaches [137, 172]. This may be particularly true for salt-bridges which involve arginine residues [172]. As such, the absence of multiple charged side chains in the sequence of PHF6 likely explains the difference between the present results and those of our prior work. For PHF6, representative structures from the lowest energy state within the explicit solvent pmf contain one salt-bridge between the side chain of lysine 6 and the C-terminal carboxyl of the same residue (Figure 12). Therefore the explicit solvent pmf suggests that the lowest energy state is extended without any salt-bridges or

hydrogen bonds between moieties that are separated in the sequence. This simple extended state which lacks salt-bridges or hydrogen bonds between distant residues is well modeled by the implicit solvent models investigated in this work.

All of the solvent models predict that PHF6 preferentially adopts extended structures in solution, and a conformational analysis of amino-acids in PHF6 argues that residues 2-5 adopt φ/ψ values corresponding to the β-strands. These findings have important implications for the pathogenesis of neurofibrillary tangle formation in patients with Alzheimer's disease. In particular, there is growing consensus that the ability of amyloidogenic proteins like tau to aggregate stems from properties of the protein backbone. In many instances, protein aggregation requires the formation of intermolecular backbone hydrogen bonds yielding a cross β-structure (i.e. the β strands are perpendicular to the axis of the fibril), and for tau this process is likely important for the initiation of neurofibrillary tangle formation ([173-175]).

Our findings imply that PHF6 exhibits a strong preference for extended β-structures in solution – a finding which suggests that PHF6 promotes neurofibrillary tangle formation by facilitating the formation of cross β-structure between tau monomers. This premise is consistent with recent data suggesting that the sequence of PHF6 is the minimal region of tau required for tau aggregation into cross-β filaments and hence neurofibrillary tangles [5]. As neurofibrillary tangle formation may play a role in neurodegeneration [114], therapies directed at modifying the structural preference for PHF6 may lead to new treatments for dementias like AD and the tauopathies [176].

Chapter 4: The Effect of a ΔK280 Mutation on the Unfolded State of a Microtubule-Binding Repeat in Tau

(This work was published as A. Huang and C. M. Stultz, "The Effect of a ΔK280 Mutation on the Unfolded State of a Microtubule Binding Repeat in Tau," PLoS Computational Biology, vol. 4(8): e1000155, pp. 1-12, 2008.)

Abstract

Tau is a natively unfolded protein that forms intracellular aggregates in the brains of patients with Alzheimer's disease. To decipher the mechanism underlying the formation of tau aggregates, we developed a novel approach for constructing models of natively unfolded proteins. The method, Energy-minima Mapping and Weighting (EMW), samples local energy minima of subsequences within a natively unfolded protein and then constructs ensembles from these energetically favorable conformations that are consistent with a given set of experimental data. A unique feature of the method is that it does not strive to generate a single ensemble that represents the unfolded state. Instead we construct a number of candidate ensembles, each of which agrees with a given set of experimental constraints, and focus our analysis on local structural features that are present in all of the independently generated ensembles. Using EMW we generated ensembles that are consistent with chemical shift measurements obtained on tau constructs. Thirty models were constructed for the second microtubule binding repeat (MTBR2) in wild-type (WT) tau and a ΔK280 mutant, which is found in some forms of frontotemporal dementia. By focusing on structural features that are preserved across all ensembles, we find that the aggregation-initiating sequence, PHF6*, prefers an extended conformation in both the

WT and ΔK280 sequences. However, in WT MTBR2 the region immediately downstream from PHF6* can adopt a loop/turn conformation while the corresponding region in the ΔK280 mutant only exhibits a conformational preference for extended conformations. As an increased preference for extended states near the C-terminus of PHF6* may facilitate the propagation of β-structure downstream from PHF6*, these results explain how a deletion at position 280 can promote the formation of tau aggregates.

Introduction

Alzheimer's Disease (AD) pathology is characterized by extracellular aggregates of Aβ-amyloid (Aβ) and intraneuronal tau aggregates, known as senile plaques and neurofibrillary tangles (NFTs), respectively [177]. Despite much focus on Aβ amyloid in AD research, tau seems to play an important role as well. For example, the number of NFTs and not the number of senile plaques in the neocortex correlates with the severity of dementia in AD patients, and there are data that imply that abnormalities in tau alone may cause neurodegeneration [12]. In light of these observations, a detailed characterization of the structure of tau protein may provide insights into the pathogenesis of AD and other neurodegenerative disorders associated with tau pathology. However, probing the structure of tau is difficult because tau protein is natively unfolded (or intrinsically disordered) in solution. Several studies suggest that tau retains its function after heat or acid-induced denaturation and both CD and x-ray scattering experiments imply that tau does not adopt a well-defined folded structure in solution [55, 178, 179]. Consequently, obtaining structural and hence functional information on tau is problematic because the direct observation of unfolded states is typically difficult to achieve experimentally.

Initially, unfolded proteins were described as random coils whose properties are derived from Flory's statistical description of chain molecules [43]. For such polymers, the radius of gyration, R_G, follows the scaling law $R_G = R_0 N^{\nu}$, where R_0 is the radius of gyration of a monomeric subunit (a function of the persistence length), N is the number of subunits in the polymer, and ν is a scaling factor that depends on the solvent characteristics. The most common measure of whether a protein behaves as a random coil is to test whether its radius of gyration follows this scaling law. However, while a structurally disordered molecule can exhibit random coil statistics, the converse is not necessarily true; i.e., random coil statistics do not imply that the structure is completely disordered [180]. Slight structural preferences may exist for some natively unfolded proteins and small changes in the distribution of conformers within an unfolded ensemble may play a role in the normal and pathological functioning of intrinsically disordered systems. A recent study, for example, suggests that inducer-mediated tau polymerization involves an allosterically regulated conformational change [181]. This is consistent with the notion that the formation of tau fibrils is associated with a shift in the conformational distribution of tau such that the unfolded state has a preference for pro-aggregatory conformations in the presence of an inducer. In light of this, constructing detailed ensembles that model the unfolded ensemble of tau may facilitate the identification of structural properties that promote aggregation.

As full-length tau contains more than 400 amino-acids (441 residues for the htau40 isoform [112]) constructing detailed ensembles that model the unfolded state of this protein is a daunting task. Fortunately, tau contains three or four imperfect microtubule-binding repeats (MTBRs) near the C-terminus of the protein and almost all known mutations of tau that are associated with inherited forms of neurodegenerative-diseases, are located in MTBR domains or their nearby flanking regions [182]. As these data suggest that MTBRs play an important role in the

progression of inherited tauopathies, we first focus on building ensembles that model the structure of individual MTBRs. It is important to note, however, that we do not strive to model the structure of a given MTBR fragment alone in solution. Rather, our goal is to generate ensembles that model the range of conformations that a MTBR can adopt when it is part of full length tau. In the present study we focus on building ensembles for the second MTBR, henceforth referred to as MTBR2. This repeat is of particular interest because it contains both a six amino-acid repeat, PHF6*, which is a minimum interaction motif that can initiate tau aggregation *in vitro* [154, 155], and the site of the pro-aggregatory mutation, ΔK280, which is associated with some forms of frontotemporal dementia [6, 7, 183, 184].

We have developed a method, called Energy-minima Mapping and Weighting (EMW), to construct ensembles that model the unfolded state of proteins. The underlying assumption that forms the basis of this approach is that the unfolded state can be modeled as a set of energetically favorable conformers, where each conformer corresponds to a local energy minimum. The method involves constructing a library of energetically favorable conformations and selecting conformations from this library to form ensembles that are consistent with a given set of experimental data. We use EMW to build ensembles for wild-type (WT) MTBR2 and the corresponding ΔK280 mutant. By comparing data from the two sets of ensembles, we deduce structural preferences in the ΔK280 ensemble that explain its increased propensity to form tau aggregates.

Results

The EMW method begins by constructing sets of energetically favorable conformations for a sequence of amino-acids within a natively unfolded protein (Figure 15). In the case of tau we

focus on MTBR2 since this region contains the aggregation-initiating sequence PHF6* as well as the site of a mutation that is associated with increased tau aggregation *in vitro* [124]. A set of local energy minima is then constructed for this subsequence, hence forming the candidate ensemble (Figure 15). Associated with each structure in this ensemble is a weight, ω_i, which corresponds to the probability that the given subsequence adopts the i^{th} conformation in the candidate ensemble. We say that an ensemble is fully specified when the local energy minima and weights are known.

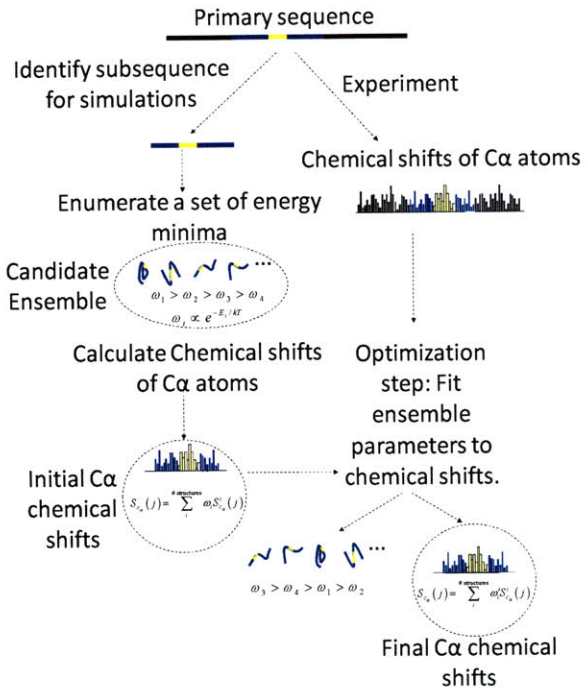


Figure 15: Outline of EMW method. The subsequence chosen for simulations is colored blue and contains an aggregation initiating sequence (colored yellow). A set of local energy minima can be enumerated using quenched molecular dynamics. Chemical shifts are calculated for the candidate ensemble and compared to chemical shifts obtained on the entire sequence. Weights of ensemble members are modified to improve agreement with experiment. $S^i_{c\alpha}(j)$ denotes the chemical shift of the Cα atom in the jth residue of the ith structure of the ensemble. $S^i_{c\alpha}(j)$ is computed from the i^{th} structure using SHIFTX [2]. $S_{C\alpha}(j)$ is the statistical mechanical equivalent of the experimentally observed chemical shift of the Cα atom in the jth residue. We note that although the aggregation-initiating sequence is shown at the center of the chosen subsequence, this need not be the case. For MTBR2, the aggregation-initiating sequence is located at the N-terminus.

Initial weights for structures in the candidate ensemble are calculated from the relative energies of each structure, as shown in Figure 15. However, as sampling is performed on a relatively small subsequence these weights may not reflect the relative probabilities of different conformations when the subsequence is part of the larger protein. For example, compact states may be preferred over extended states when the subsequence is in isolation but not when part of tau. Therefore, the composition of the ensemble is optimized and the members of the candidate ensemble are re-weighted in light of experimental data that is obtained on a larger segment of tau

protein. Sampling small subsequences increases the chance that we will observe a relatively large number of accessible states for this system. Using experimental data obtained on a larger region of tau (and not just the subsequence of interest) helps to ensure that the calculated ensemble represents the local structure of the sequence as it appears within full length tau.

A central component of EMW is that we do not strive to construct a single model for the unfolded state. We recognize that the construction of unfolded ensembles that agree with any given set of experimental data is largely an underdetermined problem; hence it is likely that there are a number of different ensembles that are consistent with a given set of experimental data. Consequently, we constructed several ensembles that are all consistent with the experimental measurements and focused our analysis on local structural motifs that are present in all ensembles. For this study, we focused on NMR data that are available for *both* WT MTBR2 and a ΔK280 mutant. These data were kindly provided by Marco Mukrasch, Daniela Fischer, and Markus Zweckstetter [123, 124].

Using the EMW method, 100 ensembles were constructed for both wild-type (WT) and ΔK280 sequences of MTBR2 (a total of 200 ensembles). Each ensemble was constructed to minimize the difference between calculated $^{13}C\alpha$ chemical shifts and the corresponding experimentally determined $^{13}C\alpha$ chemical shifts. The number of structures in each ensemble corresponds to the minimal number of structures needed to fit the available chemical shifts. Preliminary calculations found that 15 conformers were needed; i.e., fewer structures resulted in worse fits to the $^{13}C\alpha$ chemical shifts and more structures did not significantly improve the quality of fits. We note that other models examining residual structure in the unfolded state have utilized a similar number of representative conformers [57].

Application of EMW yielded ensembles that were in excellent agreement with experimentally determined absolute $^{13}C\alpha$ chemical shifts (Figure 16A & B). The average rms error between the calculated $^{13}C\alpha$ chemical shifts and the corresponding experimental values was 0.1ppm – well below the error associated with SHIFTX chemical shift predictions and similar to the error associated with experimental chemical shift measurements on K18 constructs [2, 124]. However, given that measured absolute chemical shifts for the 20 amino acids vary significantly according to the amino-acid type, reasonable correlations to absolute chemical shifts may be achieved by simply predicting amino-acid specific random coil values. Given this, we analyzed the relationship between the chemical shifts, after subtracting out residue-specific random coil chemical shift values; i.e., the secondary chemical shifts. Overall, there is excellent agreement between calculated secondary chemical shifts and the corresponding experimental values for each residue in the sequence (Figure 16C & D). These data demonstrate that the calculated models yield agreement with experiment on a per residue basis.

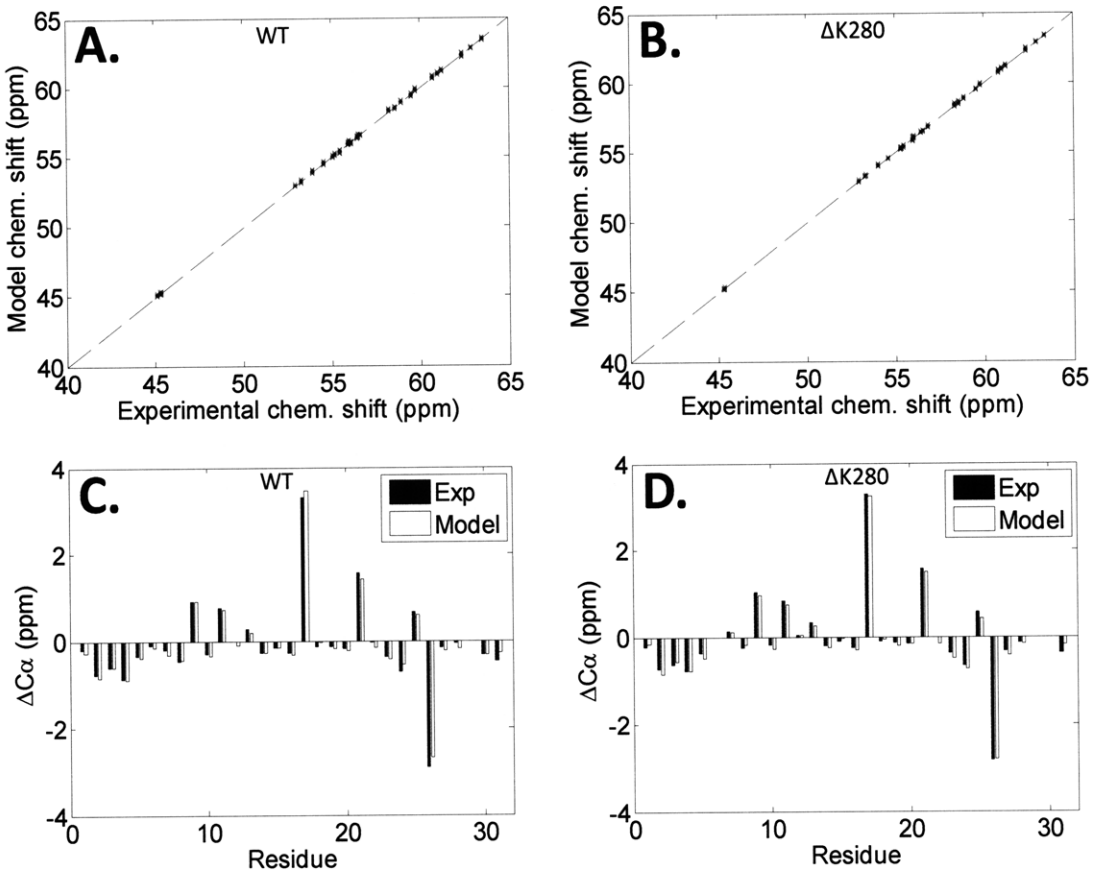


Figure 16: Model vs. experimental absolute Cα chemical shifts for A) 100 WT ensembles and B) 100 ΔK280 ensembles. Cα secondary chemical shifts (ΔCα) are also shown for the C) WT and D) ΔK280 sequences using the ensemble that had the worst agreement with experiment. The worst model is defined as the ensemble that has the greatest rms deviation between the calculated and experimentally determined values.

Ensemble	WT ^{13}CO	ΔK280 ^{13}CO	WT ^{1}HN	ΔK280 ^{1}HN
1	0.58	0.78	0.20	0.25
2	0.59	0.78	0.23	0.28
3	0.60	0.80	0.20	0.25
4	0.65	0.80	0.26	0.25
5	0.66	0.80	0.20	0.19
6	0.66	0.82	0.21	0.23
7	0.67	0.83	0.21	0.23
8	0.67	0.83	0.21	0.25
9	0.68	0.83	0.20	0.23
10	0.68	0.84	0.26	0.28
11	0.69	0.86	0.21	0.20
12	0.70	0.86	0.20	0.25
13	0.70	0.86	0.20	0.26
14	0.70	0.86	0.21	0.28
15	0.71	0.86	0.23	0.24
16	0.71	0.87	0.18	0.23
17	0.71	0.87	0.25	0.21
18	0.72	0.87	0.16	0.23
19	0.72	0.87	0.23	0.28
20	0.72	0.87	0.20	0.29
21	0.72	0.87	0.23	0.20
22	0.73	0.87	0.21	0.29
23	0.73	0.87	0.20	0.21
24	0.73	0.88	0.21	0.27
25	0.73	0.88	0.19	0.24
26	0.73	0.88	0.24	0.26
27	0.73	0.88	0.20	0.23
28	0.73	0.88	0.19	0.25
29	0.73	0.89	0.21	0.24
30	0.74	0.89	0.22	0.27

Table 3: RMS deviation between calculated and experimental CO and H chemical shifts

In the next step of our protocol, carbonyl carbon (^{13}CO) chemical shifts were used to test whether the resulting ensembles can predict experimental observations that were not used to construct the model. This helps to ensure that our models are not "overly-fit" to the $^{13}C\alpha$ chemical shifts. In general, a model that is over-fit to a given set of experimental data can reproduce that data remarkably well, but cannot reproduce data that was not used to generate the model. Therefore we consider an ensemble to be validated if new experimental results can be

accurately predicted from the ensemble. For both the WT and ΔK280 sequences, each of the 100 ensembles was ranked based on its ability to predict ^{13}CO chemical shifts. Based on these data the thirty best ensembles were chosen for further analysis. The rms difference between the calculated ^{13}CO chemical shifts and the corresponding experimental values are below 0.9ppm; i.e., below the error associated with available chemical shift prediction algorithms (

Table 3) [2]. To further demonstrate that these thirty ensembles can reproduce additional data not used in the model constructed, we computed the error between calculated amide hydrogen (^{1}HN) chemical shifts and the corresponding experimental values. The resulting values agreed with the experimentally measured ones to within 0.3ppm (

Table 3).

As expected, structures that comprise the WT (Figure 17A) and ΔK280 (Figure 17B) ensembles are heterogeneous in that they sample a wide range of conformations. Since each of the thirty ensembles represents an independent representation of the unfolded state, we searched for local structural motifs that are found in all of the ensembles. More precisely, the existence of a local conformation that is consistently adopted by a given subsequence in MTBR2 suggests that this conformation is needed to reproduce the experimental results. We therefore consider conserved motifs to represent local conformational preferences.

ure 17: A) An alignment of structures from A) all 30 WT embles. B) all 30 ΔK280 ensembles.

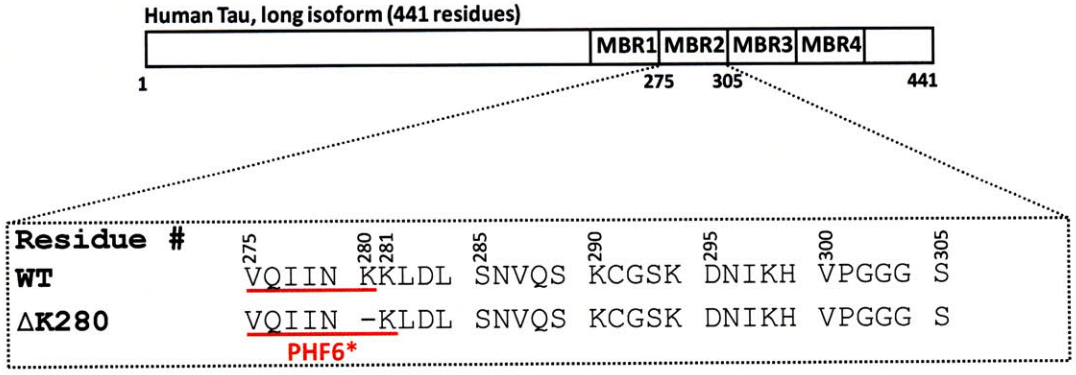
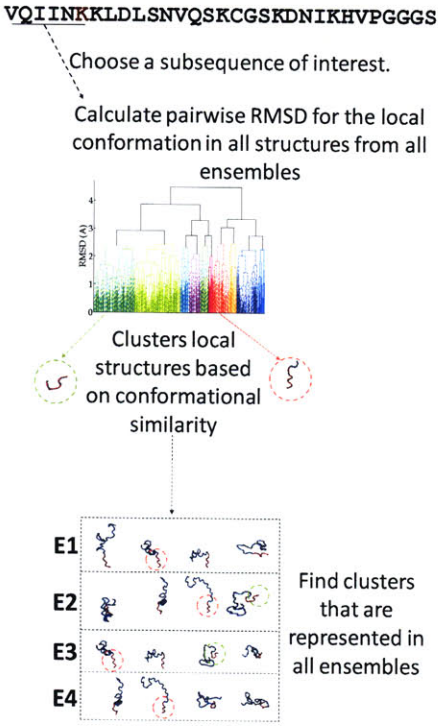


Figure 18: Sequences of WT and ΔK280 forms of the second microtubule-binding repeat in tau. The PHF6* region is underlined in red.



We begin with an assessment of the local conformation of PHF6* in both the WT and ΔK280 ensembles. Since PHF6* in the WT sequence spans residues 275-280, the ΔK280 mutant sequence has a deletion in the six-residue stretch corresponding to PHF6*. However, since residue 281 is also a lysine, the ΔK280 mutant contains an equivalent PHF6* subsequence at its N-terminus (Figure 18). This allows us to directly compare the conformation of PHF6* in both

Figure 19: Outline of the method used for clustering local conformations. First, a six-residue local region is selected for analysis. Clusters of structures with similar conformations in the region of interest are formed based on pairwise RMSDs for backbone atoms in the local region. E-E5 represent different ensembles. Clusters that are present in all model ensembles are circled in red, while unpreserved clusters are circled in green.

sequences. To identify preserved conf ormations of PHF6*, we first determined the different types of structures that this subsequence can adopt by clustering structures using only the backbone atoms of PHF6* (Figure 19). The probability that a given cluster occurs in an ensemble is equal to the sum of the weights of structures in that ensemble that contains a motif in the cluster. Preserved structural motifs are defined as clusters that have a non-zero weight in every ensemble (Figure 19); i.e., a preserved motif is found in all ensembles. For comparison, we repeated this procedure for all contiguous six-residue subsequences within MTBR2, yielding a collection of approximately 300 clusters that represent all possible structural motifs in our ensembles that any six-residue sequence in MTBR2 can adopt. Using the criterion outlined above, roughly 5% of these clusters were preserved across all ensembles.

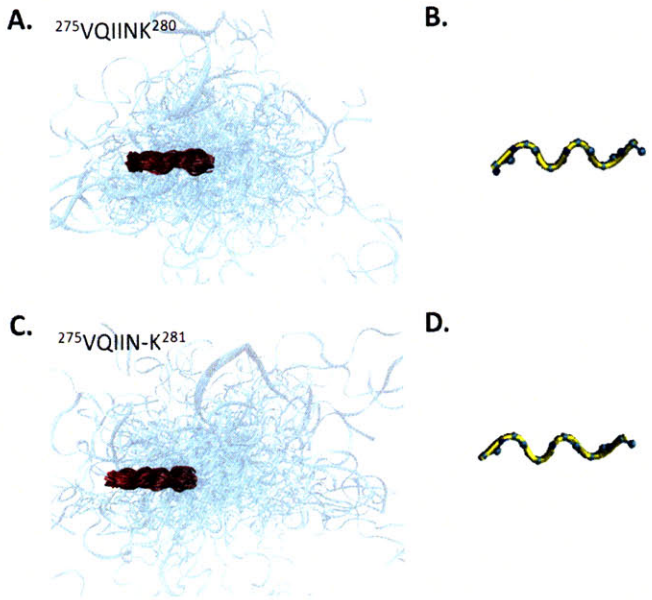


Figure 20: Structures of the cluster representing the local conformation of PHF6* that is preserved in all ensembles. A) Aligned structures for WT tau and B) Average backbone conformation for this cluster; C) Aligned structures for WT the ΔK280 mutant and D) Corresponding average structure. The backbone of PHF6* is shown in yellow for the average structures.

In WT MTBR2, clustering based on the conformation of PHF6* yielded 12 distinct conformations. However, only one of these states was present in all 30 ensembles (Figure 20A, B). Similarly, while PHF6* clusters into 11 distinct conformations in the mutant ΔK280 ensembles, only one conformation was preserved (Figure 20C & D). In both cases, the preserved conformation of PHF6* is extended conformation and has has φ, ψ angles that fall within the broad region of the Ramachandran plot corresponding to β-structure. This observation is consistent with the notion that PHF6* *a priori* adopts extended conformations that can readily form cross β-structure with other tau monomers [48]. Since the formation of cross β-structure is believed to play an essential role in the formation of protein aggregates, these data are consistent with the notion that PHF6* promotes aggregation by forming β-structure between tau monomers [154, 155].

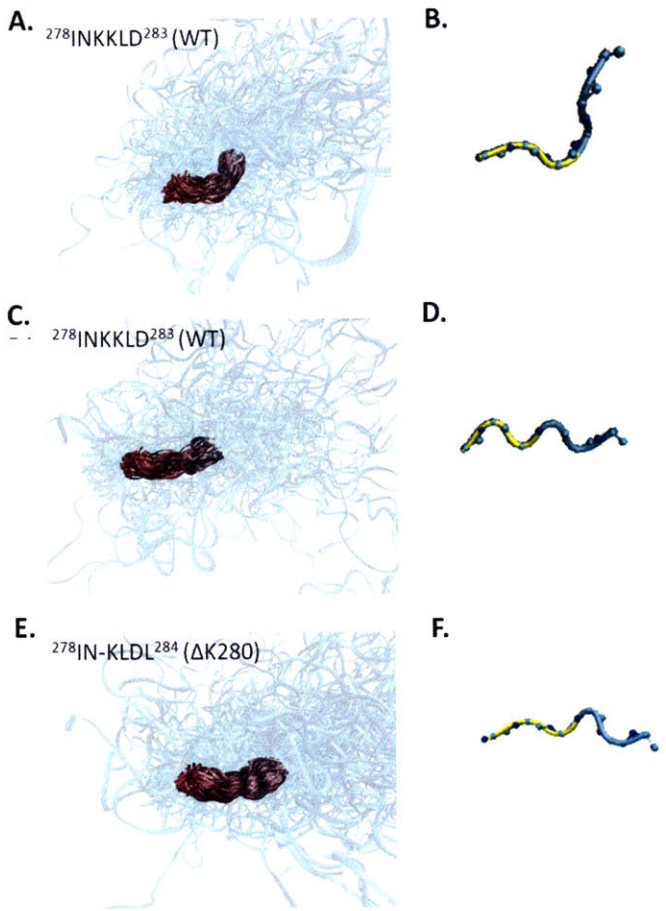


Figure 21: Preserved structures for the region corresponding to ^{278}INKKLD^{283} in both the WT (A-D) and ^{278}IN-KLDL^{284} ΔK280 ensembles (E, F). A,C) Aligned structures corresponding to a preserved cluster in the WT ensembles aligned by backbone atoms of residues of ^{278}INKKLD^{283}and B,D) the corresponding average structures. E) Aligned structures of the preserved cluster in ΔK280 ensembles, aligned by backbone atoms of residues ^{278}IN-KLDL^{284}. F) The average conformation of the preserved cluster in ΔK280 ensembles. In the average structures, residues belonging to PHF6* are in yellow.

To explore the effect of the ΔK280 mutation on the local structure of MTBR2, we analyzed the structure of the subsequences ^{278}INKKLD^{283} and ^{278}IN-KLDL^{284} in the WT and ΔK280 sequences, respectively. For WT MTBR2, two conformations for ^{278}INKKLD^{283} were found in all ensembles. The first is a loop/turn that is associated with a change in the direction of the mainchain (Figure 21A, B). In this structure residue K280 has φ, ψ angles of approximately -102 and -30, respectively; i.e., mainchain dihedral angles consistent with an α-helical/turn

conformation. The second conformation is more extended, having φ, ψ angles that place its residues within the broad region corresponding to extended β-structure (Figure 21C, D). In the mutant sequence, residue K280 is absent and the corresponding sequence, ^{278}IN-KLDL^{283}, has one preserved conformation. The deletion of residue 280, which can adopt an α-helical/turn conformation in the native sequence, results in the local region preferentially adopting an extended conformation (Figure 21E, F). The deletion, however, also introduces a slight kink in the mainchain of the sequence (Figure 21F).

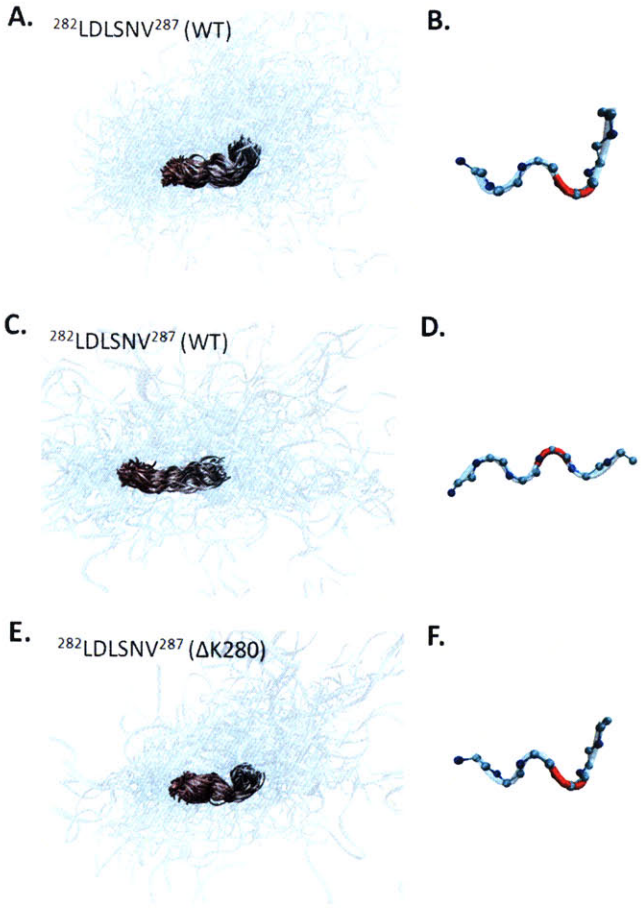


Figure 22: Preserved structures for the region corresponding to residues ^{282}LDLSNV^{287} in both the WT (A-D) and ΔK280 ensembles (E, F). A,C) Aligned structures corresponding to a preserved cluster in the WT ensembles aligned by backbone atoms of residues of ^{282}LDLSNV^{287} and B,D) the corresponding average structures. E) Aligned structures of the preserved cluster in ΔK280 ensembles, aligned by backbone atoms of residues ^{282}LDLSNV^{287}. F) The average conformation of the preserved cluster in ΔK280 ensembles. The location of S285 in the bend is indicated in red in the average structures.

In a prior work, N-H residual dipolar coupling (RDC) values were measured for residues in the WT K18 construct in polyacrylamide gel [19]. While most residues in MTBR2 have relatively large negative RDC values, S285 has a large positive value [19]. This difference can be explained by either a change in the local alignment tensor at S285, or the presence of α-helical/turn structure at this site [60, 185-187]. Accelerated molecular dynamics simulations of WT K18, however, confirm that the sequence ^{283}DLSN^{286} samples turn conformations with relatively high frequency [19]. In light of these observations, we explored the structure of the six residue segment, ^{282}LDLSNV^{287}, that includes residue S285. This region adopts two conformations that are preserved across all WT ensembles. One of the conformations contains a loop/turn (Figure 22A, B) where residue S285 has φ, ψ angles of -63, -39 respectively; i.e., near the optimal α-helical values (Figure 22B). The alternate conformation is extended and does not result in a change in the direction of the mainchain (Figure 22C, D). However, in the ΔK280 mutant, ^{282}LDLSNV^{287} has one structure that is preserved across all ensembles (Figure 22E, F). In this structure S285 again adopts φ, ψ angles (-95 and -63, respectively) that are consistent with an α-helical/turn conformation (Figure 22F). These data agree with the RDC data mentioned above and suggest that this region in both the WT and mutant sequences is able to adopt turn-like conformations in solution as well as in a polyacrylamide gel.

Discussion

Dynamical simulations provide a valuable tool for the analysis of unfolded proteins, providing insights that would be difficult to obtain from experiments alone [76]. A number of simulation methods have been developed to model the unfolded states of proteins and useful insights have been obtained with these techniques. Many of these approaches generate

ensembles by directly incorporating experimental constraints into molecular dynamics simulations in order to facilitate conformational sampling. These methods bias molecular trajectories to sample conformers that are consistent with a given set of experimental data. One problematic issue with *biased sampling*, however, is that it can suffer from over-fitting – a process which may yield a distribution of conformers that do not accurately model the range of structures that comprise the unfolded state [76]. Given this concern, a number of *unbiased* methods have been developed to generate ensembles for unfolded proteins. These approaches utilize fast algorithms, which do not employ a physical potential energy function, to obtain representative structures of the unfolded state, and in some cases experimental data can then be used to improve the resulting ensembles [68, 77, 188, 189]. The algorithm ENSEMBLE, for example, adjusts population weights for pre-generated conformers to improve agreement with experimental data in a manner similar to that described here [77].

A unique feature of the present method is that it does not strive to generate a single ensemble that represents the unfolded state. Given that accurate modeling of an unfolded protein is an undetermined problem, it is likely that there are a number of different ensembles that agree with any given set of experimental data. Moreover, given the immense number of potential conformations that an unfolded protein can adopt, this may be true even when a relatively large number of experimental constraints are used to construct the ensemble. Hence our goal was to construct several candidate ensembles, each of which agrees with a given set of experimental constraints, and focus our analysis on local structural features that are preserved across all ensembles. Local structural features that are found in all independent ensembles likely represent motifs that are required to reproduce the experimental data. In other words, given the underdetermined nature of the problem, it is not clear how to determine when one has the

"correct" ensemble. However, local structural motifs that consistently appear in all independent ensembles are likely to also be present in the "correct" ensemble. Consequently, we consider locally preserved structural motifs to represent local conformational preferences.

An important consideration in our method is the choice of experimental data that is used to build and validate the constructed ensembles. In principle, EMW can use any set of experimental measurements to optimize and validate model ensembles. Indeed, as more structural information is made available, additional data can and should be used to further refine the set of model ensembles. In this regard, we note that although a number of NMR measurements have been made on native tau constructs, the data available for constructs containing a ΔK280 mutation is relatively limited. In a prior study, nuclear chemical shifts and HSQC spectra were measured for the K18ΔK280 construct, which contains all four MTBRs and the ΔK280 mutation [124]. Data were obtained for both free K18ΔK280 and for K18ΔK280 in the presence of the polyanion heparin and microtubules [124]. However, as we are interested in building structural models for MTBR2 in solutions free of compounds that promote tau self-association (e.g., heparin) and free of proteins known to bind tau, we focused on measurements obtained with the free K18ΔK280 construct. Additionally, as there are a number of existing methods that relate chemical shift measurements to three dimensional protein structures [2, 190-192] we considered $^{13}C\alpha$, ^{13}CO, ^{1}HN and ^{15}N chemical shift measurements; i.e., the only available chemical shifts for K18ΔK280 [124]. Futhermore, established methods for estimating NMR chemical shifts can predict carbon and amide proton chemical shifts with an error of approximately 1ppm or less, while the error associated with predicting nitrogen chemical shifts is substantially larger (~2-2.5ppm) [2, 191-193]. Therefore we focused on the $^{13}C\alpha$, ^{13}CO and ^{1}H chemical shifts for this study because these data represent measurements that can be calculated

with the greatest accuracy and that are available for both native tau constructs and the ΔK280 mutant.

It has long been recognized that chemical shifts of a given residue are, in general, largely a function of the local environment of the residue in question [194, 195]. Since we generate ensembles that agree with chemical shifts, a limitation of the results reported here is that we do not explicitly include experimental data that more directly reveals information about non-local interactions. While long range contacts have been identified in some natively unfolded proteins (e.g., [57]), the dimensional scaling characteristics of intrinsically disordered proteins suggests that stable long-range contacts are sparse in these systems [196]. Nevertheless, we suggest that the combination of a physical potential energy function, which can in principle model long range interactions, and experimentally determined chemical shifts can provide insight into the structure of proteins in general. In this regard we note that data are emerging which suggest that backbone chemical shifts, when used in conjunction with a physical energy function, may be sufficient to adequately predict tertiary folds, and consequently stable non-local contacts, for some proteins [197, 198].

Although our work focuses on the structure of the MTBR2 without explicitly including other MTBRs, our findings may also have implications for full length tau. Once a representative set of conformers for MTBR2 is generated, we strive to ensure that the calculated chemical shifts agree with chemical shifts obtained using a construct that contains all MTBRs. This helps to guarantee that the ensemble models the structure of MTBR2 as it appears in full length tau. In short, we are not interested in the structure of MTR2 as it appears alone in solution; instead we hope to deduce structural features of MTBR2 as it appears in full length tau. In addition, as MTBR2 contains an aggregation-initiating sequence that is known promote tau aggregation *in vitro* as

well as the site of a mutation that leads to in increased tau aggregation *in vitro* and *in vivo*, studies of both its WT and mutant forms may lead to insights into the mechanism of tau aggregation [6, 155, 199].

The ability to form intermolecular β-sheet conformations appears to be a relatively general property of polypeptide chains that are associated with disorders of protein misfolding and aggregation [36, 173, 200, 201]. Therefore it is likely that an inherent propensity to form extended conformations, that are consistent with β-structure, will promote aggregation in natively unfolded systems. When EMW is applied to MTBR2, we find that the aggregation-initiating sequence, PHF6*, adopts an extended conformation in both the WT and ΔK280 ensembles, a finding consistent with the observation that these peptides can initiate tau aggregation [154, 155]. Interestingly, in a prior work we demonstrated that a related hexapeptide, PHF6, preferentially adopts an extended state that can facilitate the formation of cross- β-structure between tau monomers [48]. The present study suggests that this property is preserved when aggregation-initiating sequences are part of their corresponding MTBRs. That is, PHF6* *a priori* adopts extended conformations that can readily form hydrogen-bonded β-structure. Additionally, a recent survey of amyloidogenic proteins suggests that fibrillogenesis for natively unfolded proteins involve the formation of partially folded intermediates that can subsequently go on to form amyloid fibrils [36]. Our findings are consistent with these observations in that our results imply that formation of a locally stable, and extended, conformation plays a role in the formation of tau aggregates.

Recently, several studies have attempted to characterize residual structure of MTBRs in tau [19, 122-124, 202, 203]. These studies can be roughly divided into two categories: descriptions of ensemble average characteristics based on NMR measurements [19, 122-124], and NMR

solution structures of local regions obtained by adding organic solvents to stabilize a unique fold [202, 203]. Since the presence of organic solvents leads to significant changes in the conformational distribution of states, as evidenced by the dramatic changes in the CD spectra [55, 202, 203], the physiologic relevance of these latter results remains unclear. However, early characterizations of MTBRs in non-organic solvents, found that the PHF6 region likely has a higher propensity for extended, β-strand like conformations – a finding in accord with our data [122, 123].

Given that both WT and ΔK280 tau contain aggregation-initiating sequences (Figure 18), it is not clear how β-strand propensity in this region explains the difference in aggregation potential between the two sequences. Therefore to deduce structural features of the ΔK280 mutant that explain its proclivity to form aggregates, we analyzed the structure of MTBR2 in the vicinity of the mutation site. Unfolded ensembles of WT MTBR2 contain two conformations at the mutation site that were present in all ensembles – a loop/turn conformation and an extended state. In contrast to the WT MTBR2 ensembles, models of ΔK280 in the same region had one conformation that was present in all ensembles. This state is relatively extended and contains a kink at the site of the deletion. While the slight disruption in the extended state of the mutant may also influence the ability to form hydrogen-bonded cross-β structure, a loop/turn at the C-terminus constitutes a much greater impediment to the formation of β-structure. Since residue K280 has a relative preference for non-extended states, deletion of this residue leads to increased sampling of extended states downstream from PHF6*. The relative preference for extended structures downstream from PHF6* in the ΔK280 mutant suggests that the ability to propagate β-structure distal to PHF6* can affect the aggregation potential of tau. These observations therefore explain how the deletion of a single residue can change the aggregation potential of tau.

We also find that in both WT and mutant ensembles residue S285 can adopt φ, ψ angles consistent with an α-helical/turn structure. Recent data on the WT sequence are also consistent with these observations as RDC values and molecular dynamics simulations suggest that S285 adopts an α-helical/turn structure. Since those experiments were preformed in polyacrylamide gel, our data suggest that this structure also occurs with relatively high frequency in solution. It is also worthwhile to note that although we find that a six-residue region including K280 can adopt a similar loop/turn conformation, the associated RDCs for this region are not associated with a change in sign, like that observed at S285 [27]. Nonetheless, unlike RDC measurements for folded proteins, RDC values for unfolded proteins can be difficult to interpret [204]. This is due, in part, to the fact that prior to the measurement of RDC values, the protein of interest must first be embedded in an alignment medium [60]. This induced steric alignment of unfolded proteins may lead to results that do not fully capture the range of structures that an unfolded protein can adopt in solution. Hence the absence of particular RDC values in polyacryalmide gel (or any other alignment media) does not necessarily imply that a given conformation is not present in solutions containing the unfolded protein of interest.

The formation of tau aggregates is likely a complex process as a number of factors have been shown to influence the formation of tau aggregates *in vitro* [12, 177, 178]. Consequently, there may be additional factors that contribute to the increased ability of the ΔK280 mutant to form aggregates; e.g., a ΔK280 mutation leads to an overall decrease in the strength of the intermolecular charge-charge repulsion between tau monomers that self-associate [155]. Nonetheless, our data demonstrate that small changes in the sequence of tau can lead to localized structural changes in the unfolded ensemble that may affect tau's ability to form cross β-structure. Overall, our data suggest that small sequence-specific changes can promote tau

aggregation and that interventions that prevent the propagation of β-structure downstream from aggregation-initiating sequences, may form the basis for therapies that prevent tau aggregation.

Methods

Energy-minima Mapping and Weighting

The EMW method constructs ensembles for unfolded proteins that are consistent with a given set of experimental data. Our model for an unfolded ensemble consists of structures corresponding to local energy minima and associated probabilities (weights) that are assigned to the different conformations. For this work, the experimental measurement used to optimize and validate the model ensembles are chemical shifts for the second tau microtubule binding repeat [124]. In principle, EMW can be used with any given set of experimental data. In this application we focus on chemical shifts that were available for both the K18 and K18ΔK280 constructs

The EMW method can be decomposed into three steps i) conformational sampling, ii) model optimization, and iii) ensemble validation. Conformational sampling uses high temperature molecular dynamics (MD) followed by minimization of the resulting structures (i.e. quenched dynamics) to create a library of widely varying conformations representing minima on the potential energy surface. Model optimization is performed to select a subset of these structures and optimize weights that represent the relative prevalence of each structure. Validation is performed by computing additional chemical shifts that not used to construct the ensemble and comparing these data to experimentally measured carbonyl carbon shifts. In what follows we outline each step of the EMW method.

Conformational Sampling

We used quenched molecular dynamics (QMD) to sample different local energy minima of the R2 peptide. Conformational sampling was performed on a blocked peptide with the sequence corresponding to the second microtubule binding repeat. A polar-hydrogen model of the WT (VQIINKKLDLSNVQSKCGSKDNIKHVPGGGS) and ΔK280 (VQIINKLDLSNVQSKCGSKDNIKHVPGGGS) MTBR2 peptides were constructed using CHARMM [157]. The N and C-termini were blocked using ACE and CBX residues defined in the Effective-Energy Function-1 (EEF1) model [140]. This sampling procedure consisted of high temperature molecular dynamics (used to randomize the initial conformation of the protein) followed by quenched dynamics. To ensure that a wide range of conformations was sampled, constraints were imposed on the peptide for the high temperature and quenching steps. Specifically, conformational sampling was performed in a series of molecular dynamics simulations. In each simulation the end-to-end distance of MTBR2 was restrained to a pre-defined value; i.e., 3Å, 4Å, 5Å , ..., 70Å, where the end-to-end distance was defined as the distance between the Cα carbons on residue VAL1 and SER31 of the peptide. End-to-end restraints were used to ensure that both compact and extended states were sampled during the high temperature simulations. For each end-to-end distance, 4ns of high temperature MD at 1000K was performed with the EEF1 implicit model of solvent [140]. All simulations employed a Berendsen thermostat to maintain the system temperature at the desired value [159]. Hydrogen bond lengths were held near their equilibrium values using SHAKE [160] and a 2 fs timestep was used. Coordinates were saved every 10ps, yielding a total of 400 structures per end-

to-end distance. This procedure was applied to both WT and ΔK280 sequences, producing a total of 27,200 structures for each sequence.

Each structure was then used to initiate a new MD trajectory which cools the system to 298K over 40ps of simulation by coupling the sampled system (including atom coordinates and corresponding velocities) to a Berendsen heat bath at 298K. At the end of this cooling simulation, structures were minimized for 10000 steps using the Adopted Basis Newton Rhapson algorithm [157]. Restraints were removed for the minimization step to ensure that minima on the unbiased energy surface are sampled. Searching for minima in the vicinity of the randomized conformation by cooling and equilibration followed by minimization rather than simply performing direct minimization allows the structures to escape shallow local energy minima and find more stable states.

As the conformation of PHF6* is of particular importance, additional simulations were performed to ensure that a large range of PHF6* conformations were represented in the ensembles. Each additional simulation constrained the PHF6* radius of gyration to adopt a pre-defined radius of gyration (4Å -5.9Å) while the restricting the end-to-end distance of MTBR2 to be near 9Å. This was done because our initial data suggested that compact conformations of MTBR2 were relatively undersampled after early QMD simulations. In total 31,200 local energy minima were generated for the native polypeptide and 31,200 structures were generated for the mutant structure. We refer to this set as our *structure library*.

We note that no single structure in our structure library had calculated backbone chemical shifts that agreed with the corresponding experimental values. For example, amongst the 31,200 structures, we found one conformer that had a $^{13}C\alpha$ chemical shift error of approximately 1ppm (compared to the ensemble shift errors of 0.1ppm). In addition, this structure had a ^{13}CO

chemical shift error of 2.3ppm (compared to the ensemble CO errors which were all below 0.9ppm).

Ensemble Optimization

The optimization procedure strives to obtain ensembles that have calculated chemical shifts that agree with experiment. The function to be minimized is:

$$f\left(\{\omega_i, X_i\}_{i=1}^{N}\right) = \sum_{j=1}^{r}\left(S_{c_\alpha}(j) - S_{c_\alpha}^{Exp}(j)\right)^2 \tag{4.1}$$

where N is the number of structures in the ensemble, X_i is the Cartesian coordinates of the i^{th} structure, ω_i is the weight of the i^{th} structure, r is the number of residues in MTBR2, $S_{c_\alpha}^{Exp}(j)$ is the experimentally determined $C\alpha$ chemical shift of residue j, and $S_{c_\alpha}(j)$ is the calculated $C\alpha$ chemical shift of residue j. Using the definition of $S_{c_\alpha}(j)$ shown in Figure 15 we have:

$$f\left(\{\omega_i, X_i\}_{i=1}^{N}\right) = \sum_{j=1}^{r}\left(\sum_{i}^{N}\left(\omega_i S_{c_\alpha}^{X_i}(j)\right) - S_{c_\alpha}^{Exp}(j)\right)^2 \tag{4.2}$$

where $S_{c_\alpha}^{X_i}(j)$ is the calculated chemical shift of residue j in structure X_i. $S_{c\alpha}^{X_i}(j)$ is computed using SHIFTX [2]. We note that reported errors for the experimentally determined chemical backbone shifts are all approximately 0.1ppm [124]. Therefore, the experimental errors of individual shifts are not explicitly included in equation (4.2). Lastly, errors reported in the text represent $\sqrt{f}$ and are therefore in units of parts-per-million (ppm); i.e., the same units used for chemical shift data.

We used a simulated annealing algorithm to minimize f in equation (4.2). To implement a simulated annealing protocol we first need an initial ensemble. The candidate ensemble was constructed by dividing the structure library into n different sets based on the radius of gyration

of the different conformers, (n was allowed to vary between 1 and >100, see below). One structure was randomly chosen from each set to form the initial ensemble. This ensures that our simulated annealing protocol begins with a set of structures that span many different radii of gyration for the molecule. The weights for structures in this ensemble were calculated from the relative energy of each conformation as follows:

$$\omega_i = \frac{e^{-\frac{(E_i - TS_i)}{kT}}}{\sum_j e^{-\frac{(E_j - TS_j)}{kT}}} \tag{4.3}$$

where the energy associated with each conformation, E_i, is the EEF1 potential energy, S_i is the vibrational entropy, and T=298K [48]. This initial model (structures and weights) was the starting point of our simulated annealing protocol.

In our simulated annealing protocol, one performs a number of Monte Carlo steps at a given value of a control parameter (also referred to as the temperature). As the control parameter is gradually decreased, the system approaches its global minimum [205]. Central to any simulated annealing method is the protocol for decreasing the control parameter; i.e., the cooling schedule. We use a cooling scheduled based on the work of Nulton et al. and described in reference [206, 207].

Each Monte Carlo step consisted of several stages:

- Generating a new candidate ensemble:

  At each MC step, a structure from the current ensemble was replaced by a new structure from the library of minima (structure library) sampled by QMD to create a new candidate ensemble.

- Choosing weights for a given set of structures:

Given a new choice of 15 structures, weights were optimized using an minimization algorithm that employs an interior-reflective Newton method, to find a set of weights, ω_i, which minimize [208, 209] equation 2.

- Metropolis Acceptance Criteria

The new ensemble (structures and weights) is accepted or rejected based on a Metropolis criterion.

The simulated annealing algorithm was implemented MATLAB (© Mathworks). The number of Monte Carlo steps for a given value of the control parameter is as described in a previous work [206].

To determine the appropriate number of conformers in each ensemble, we performed the optimization procedure described above assuming that the ensemble had *n* structures, where n ranged from 1 to >100. These calculations found that a minimum of approximately 15 conformers were needed to fit the Cα chemical shifts to within 0.1ppm, which is approximately equal to the experimental error associated with these chemical shift measurements [124] and well-below the error associated with SHIFTX chemical shift predictions [2]. Including additional structures did not significantly improve the error.

Ensemble Validation

Validation consists of computing chemical shifts, using the final optimized model from, and comparing these data to experimentally measured values that were not sued in step (ii). As described in the text, ^{13}Cα-chemical shifts were used to construct the model and ^{13}CO and ^{1}HN shifts were used for validation purposes. The error between calculated and measured shifts is computed using equation (4.2), with ^{13}CO atoms substituted for ^{13}Cα atoms. Models were

ranked by their error and the 30 models with the best agreement with the ^{13}CO shifts were selected for more detailed analysis as described in the text. To further test whether these models could be used to calculated quantities not used in model construction we computed ^{1}HN chemical shifts from these thirty ensembles and compared these data to the corresponding experimental values.

Identifying Locally Preserved Conformations

We searched for conformations of 6-residue subsequences that are present in every ensemble. Six residues was a natural characteristic size for a local region of interest, as it is the length of PHF6*. To this end, all structures in each ensemble of either WT or ΔK280 MTBR2 were clustered using a matrix consisting of the pairwise rmsd backbone deviation of the each contiguous six-residue segment. Structures were clustered using MATLAB (© Mathworks) such that the maximum RMSD between two structures in a cluster was 2.5Å. A range of maximum RMSD values (1-6 Å) were examined empirically, and it was found that a cutoff of 2.5Å was sufficient to prevent similar conformations from being divided into separate clusters, while also ensuring that clusters included a relatively homogeneous set of conformations. The probability that a given cluster occurs in an ensemble is equal to the sum of the weights of all structures that contain that motif. Preserved local structural motifs were found by identifying clusters where the total weight of its structures was non-zero across all ensembles.

Structures for each cluster were visualized in VMD. To facilitate visualization of the overall conformation associated with a cluster, an average structure for each cluster was generated after 5000 steps of steepest descent minimization to remove bad contacts (only the 6 residues were minimized). Visual inspection verified that the energy minimized structures did not differ

significantly from their un-minimized counterparts. All molecular structures were made with VMD [163].

Acknowledgements

We would like to thank Marco D. Mukrasch, Daniela Fischer, and Markus Zweckstetter for providing chemical shift values from [123, 124].

Chapter 5: Models of K18

Introduction

Recently, residual dipolar couplings and radius of gyration measurements for the microtubule-binding repeat domain of tau have been published [18, 19]. The microtubule-binding repeat domain consists of the 130 residue region of tau encompassing the four microtubule-binding repeats, MBR1 through MBR4, in Figure 18. The construct for the polypeptide corresponding to the microtubule-binding repeat sequence has been designated K18 [19].

We first attempt an alternate modeling approach - to construct a model of the unfolded state of K18 without utilizing experimental data, based on physical principles. However, we found that this model was unable to fully reproduce experimental measurements. In particular, agreement with residual dipolar coupling measurements could not be obtained. Thus, in order to incorporate these K18 measurements into our analysis, we applied the previously described EMW method to obtain and analyze ensembles incorporating the new data.

The Segment Model

In principle, one could efficiently sample unfolded state conformations from first principles given a complete understanding of the physical interactions involved. At the present, the intramolecular and solvent interactions which determine conformational preferences in the unfolded state are not well characterized. Most potential functions are tested and validated on folded states of proteins [138-140]. Ideally, these potentials should generalize to either folded or

unfolded states of proteins. In practice, simplifying approximations, such as the absence of atom polarizability, may limit the physical contexts to which these potentials can be generalized. Thus, in the EMW method, we combined conformational sampling using a physical potential surface with an empirical selection and weighting of conformations to construct conformational ensembles [49]. With this approach, there exists a degeneracy of solutions – more than one set of structures can be consistent with the available experimental data. Since different solutions can exhibit different conformational properties, the conclusions one can draw from such models are limited. In this chapter, we attempt to efficiently sample conformations of K18 without fitting to experiment by proposing a hypothesis regarding the physical properties of the unfolded state.

Given the scarcity of stable long-range contacts in intrinsically disordered proteins, we hypothesized that the distribution of conformational preferences could be described by a model which accounts for local interactions but approximates sequentially-distant conformations as independent. If this hypothesis is correct, one can sample conformations of the unfolded state by utilizing distributions of local conformations from simulations of isolated sequence segments (Figure 23A (1)). These isolated segments should include overlapping residues. For example, in our implementation, we chose 8 residue segments with 3 residues of overlap (see methods) - the first segment corresponds to residues 1-8 in the sequence, the second segment corresponds to 5-12, the third segment corresponds to 9-16, etc. Once local segment conformational distributions are generated, structures of the larger protein can be efficiently generated by successively sampling conformations from these distributions. Each segment is aligned to the adjacent segments using the overlapping residues (Figure 23A (2)). Duplicate atoms are removed and the coordinates for the segments are merged into a single polypeptide chain which is minimized to remove bad contacts (Figure 23A(3)). Figure 23B shows this process of chain elongation in three

dimensions using atomic coordinates sampled from local conformational distributions. We refer to this method as the segment model. Further implementation details of this method are discussed in the methods.

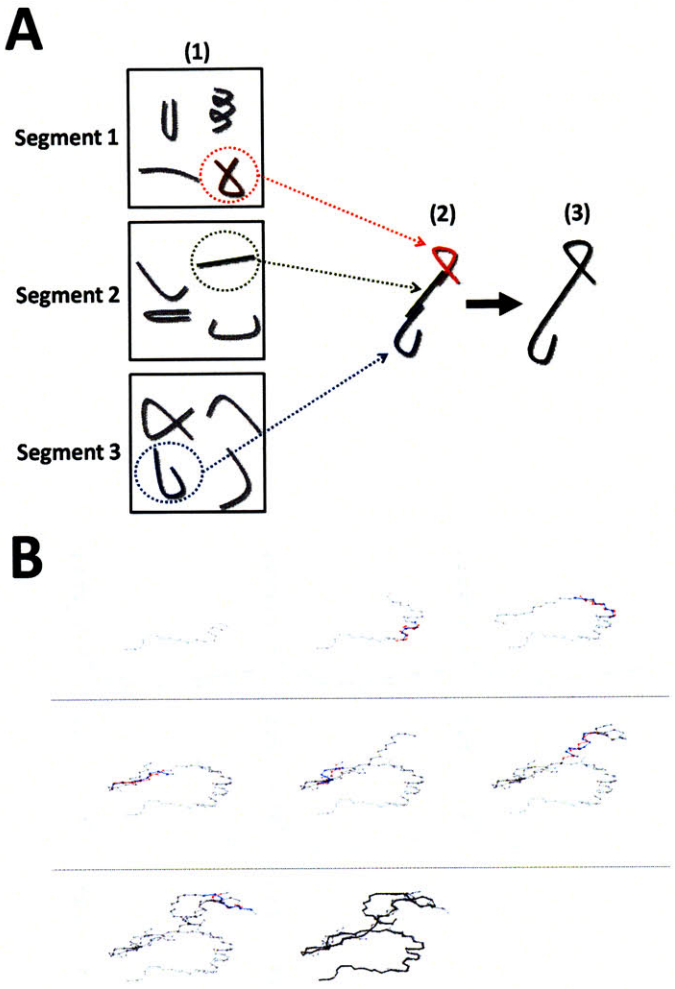


Figure 23 Sampling an unfolded protein conformation from peptide segments. (A) Schematic of the method for building structures for a polypeptide from segments. 1. Replica exchange simulations on peptide subsequences are performed to construct distributions of segment conformations for overlapping subsequences of the protein. 2. For each segment distribution from step 1, a single conformation is randomly sampled and overlapping residues are aligned. 3. Duplicate atoms are removed and atoms from segments are joined into a single polypeptide chain representing one realization of the protein conformation in the unfolded state. The final structure is minimized to remove bad contacts. This process is repeated to sample different conformations of the protein (B) The process of chain elongation shown in 3D, using atom coordinates from peptide simulations. From left-to-right, top-to-bottom, each structure shows the alignment and addition of a single local segment conformation to the overall chain. The bold line in the final structure represents the combined chain representing the larger unfolded protein.

The implication of this hypothesis is that if one can fully specify the conformational distributions of local segments in the intrinsically disordered protein, then the conformational distribution of the larger protein is fully specified. Thus a full characterization of the conformational space does not require description of a conformational distribution of size $O(s^N)$ (N is the number of residues in the protein of interest, s is the number of states available to each residue), but rather $O((N/n)s^n)$ conformations describing the set of local conformational distributions (n is the number of residues in the subsequence). In general $n<N$ and therefore $(N/n)s^n << s^N$. If such a model is adequate, the complexity associated with describing the unfolded ensemble becomes far more tractable.

Statistical coil models (see page 21) can be thought of as a special case of the segment model in which n=2 (Figure 24). Likewise an all-atom, full protein simulation can be thought of as a special case of the segment model in which n=N. If one considers a spectrum of models for intrinsically disordered proteins, the segment model corresponds to an intermediate degree of detail (for $2<n<N$) between statistical coil models and full protein simulations.

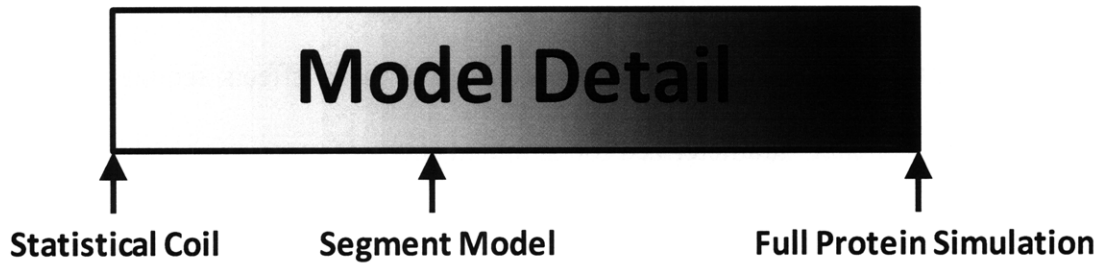


Figure 24 The segment model represents an intermediate level of detail for modeling intrachain interactions. Statistical coil models generally model nearest neighbor effects (and volume exclusion), while all-atom MD models include interactions between all atoms.

The segment model differs from the statistical coil model in that it can account for interactions (other than volume exclusion) and correlated conformations within a segment. Furthermore, in our implementation, local conformational preferences are derived from a

physical potential function modeling physiological conditions, rather than phi/psi propensities tuned to reproduce RDC measurements under denaturing conditions [67, 68].

Results

The Segment Model

Models of K18 were constructed by sampling and connecting peptide conformations as described in the methods section. 5000 structures were sampled (a comparable number of structures are used in other stochastic models of the unfolded state [67, 68]). We began evaluating the ensemble by comparing the ensemble-average radius of gyration to measured values obtained by SAXS [18]. The segment model substantially underestimates the average radius of gyration of the ensemble, computing a radius of gyration of 1.81nm, whereas the measured RGYR of K18 is 3.8±0.3nm.

One explanation for this discrepancy is that intramolecular electrostatic interactions in the implicit solvent potential are overemphasized, biasing the distribution of peptide conformers in favor of compact structures [136]. Fully compensating for these effects requires an implicit solvent potential that is better parameterized for unfolded states of proteins, which is beyond the scope of this work. Another contributing factor to this discrepancy is that volume exclusion due to the remainder of the protein is not accounted for in generating local conformational distributions.

We examined whether experimental quantities could be reproduced by utilizing the existing conformational distributions and sampling structures with a modified probabilitydistribution which favors selection of extended conformations during the construction of K18 structures:

$$P(s_{i,j}) = \frac{e^{-\lambda(Rg_{i,j}-Rg_E)^2}}{\sum_k e^{-\lambda(Rg_{i,k}-Rg_E)^2}} \quad (5.1)$$

where $P(s_{i,j})$ is the probability of sampling peptide structure j at segment i when constructing a K18 structure, $Rg_{i,j}$ is the backbone radius of gyration of peptide structure j at segment i, k iterates through all structures at segment i, Rg_E is the backbone radius of gyration of a fully-extended eight-residue peptide (8.5Å), and λ is the scaling parameter for favoring extended conformers. This formalism is equivalent to introducing a harmonic potential that is centered at the fully extended state with λ as a force constant. For $\lambda = 0$, this distribution reproduces the uniform sampling of conformers from the REMD simulation. By biasing the local conformational distributions towards more extended conformations, the distribution of the generated K18 structures becomes more extended as well (Figure 25). One limitation of this method is that the sampling of segment conformations is restricted to the existing sampled conformations. Furthermore, since the free parameter, λ, is fit to experiment, experimental measures other than radius of gyration (i.e. chemical shifts and RDCs) must be utilized to validate the correctness of the resulting ensemble.

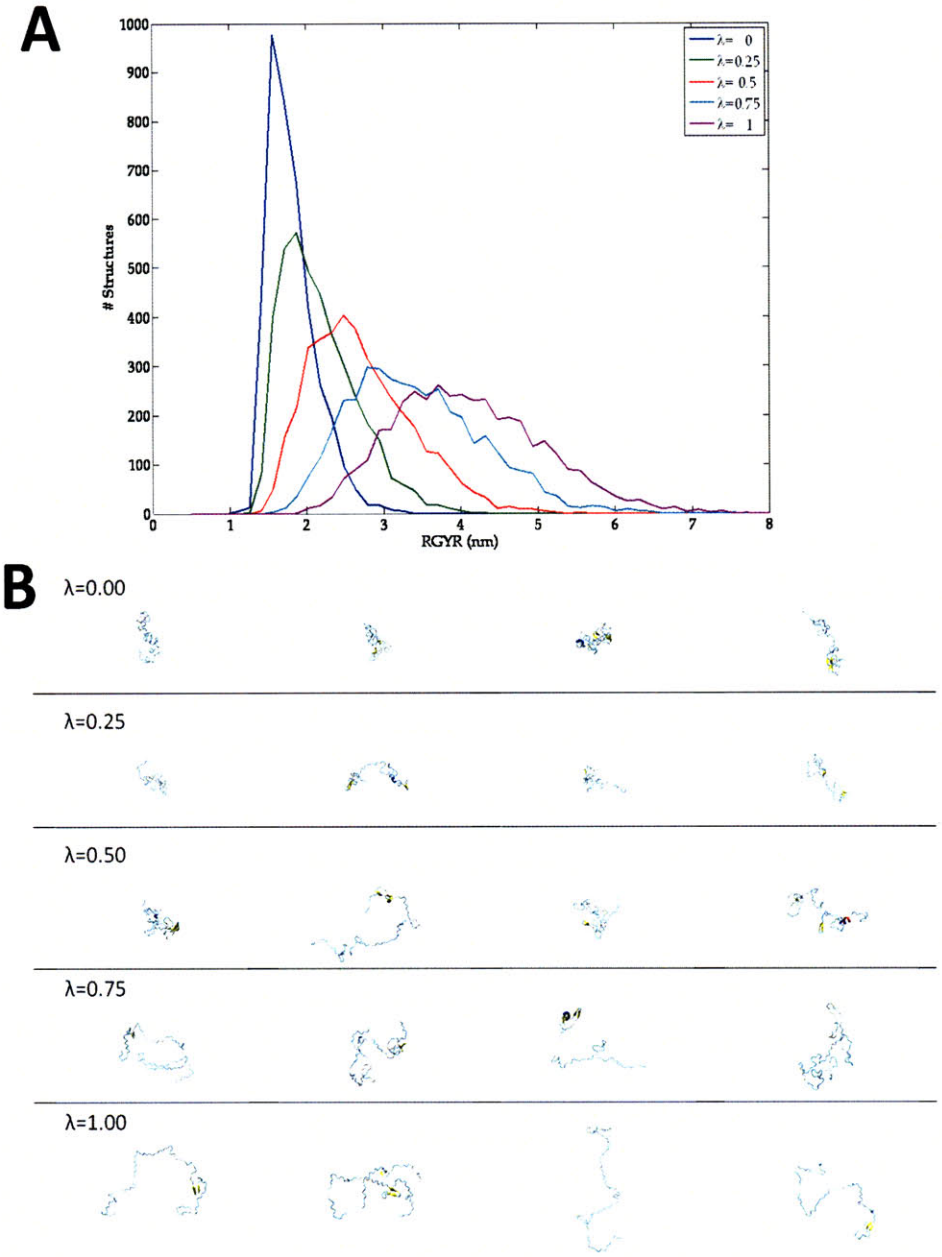


Figure 25 The effect of biasing sampling of segments towards more extended conformations. (A) The distribution of K18 RGYR as λ is varied from 0 to 1. (B) Example K18 conformations as λ is varied from 0 to 1.

A parameter value of $\lambda = 0.875$ resulted in an ensemble with an average radius of gyration equal to the experimental measurement of 3.8nm. We proceeded to compare this ensemble with chemical shift and RDC measurements. For each of the structures in the ensemble, chemical

shifts were computed with SHIFTX and RDC values were computed with PALES and the ensemble average calculated [2, 82]. Consistent with previous studies, ensemble averaged RDCs were scaled by a constant scaling factor for comparison with experiment, as has been utilized in previous models to equalize the range of simulated and experimental RDCs [67, 68].

Comparisons between the ensemble average H, Cα, and CO chemical shifts and experimental measurements yielded reasonable agreement with root mean squared errors of 0.73ppm, 0.23ppm, and 0.77ppm, respectively - all within the prediction error of SHIFTX [78]. Agreement with nitrogen chemical shifts was poor, with an error of 3.96ppm, however the prediction error for nitrogen chemical shifts is substantially larger than for other atom types and the relationship between nitrogen chemical shifts and protein structure is less-well understood [2]. The incorporation of sidechain coordinates from the segment conformations into the K18 structures is required for agreement with experiment. If sidechains are built and minimized from scratch after construction of the K18 backbone instead of incorporated from the peptide coordinates, chemical shift agreement is worsened, particularly for CO chemical shifts whose errors increase by 62%. However ensemble average RDCs of the segment model do not exhibit agreement with experiment (not shown). Possible reasons for this are examined in the discussion section. We concluded that the current implementation of the segment model was insufficient to reproduce the experimental properties of K18.

Energy Minima Mapping and Weighting Models of K18

We examined whether EMW could be used to select and weight structures to generate models of K18 consistent with experiment [17, 49]. Since RDCs proved to be the key difficulty in reproducing experimental measurements for the segment model, they were utilized in the

EMW optimization rather than chemical shifts. The previously described EMW protocol [49] was adapted for RDCs (see methods for details) and 70 ensembles were generated.

Resulting ensembles were observed to be consistent with experimental Cα, CO, and H chemical shifts to within prediction error. RDCs of these models were substantially improved by the fitting. In contrast to the segment model, this procedure was able to show qualitative agreement with measurements (Figure 26), though future work may utilize larger ensembles to further improve the agreement with measured RDC values. For comparison, we generated EMW models of ΔK280 forms of K18 using available experimental data.

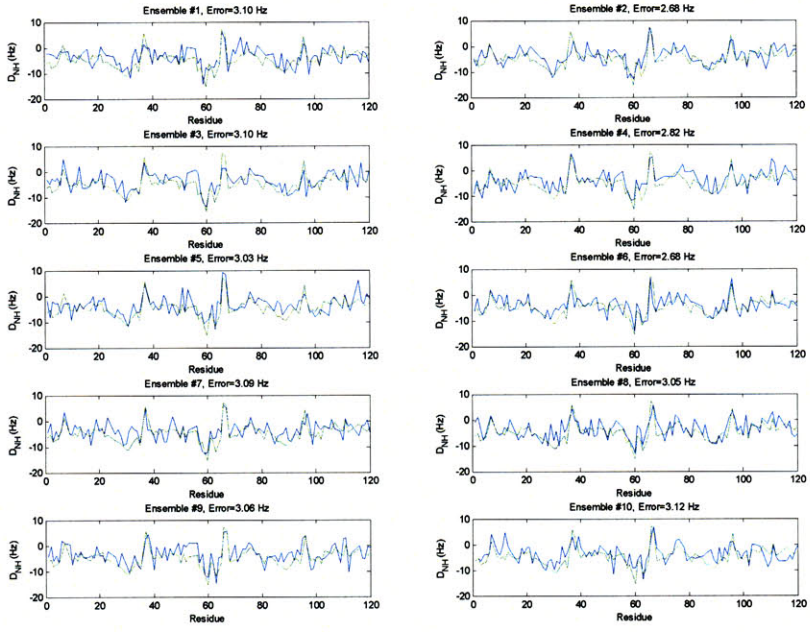


Figure 26 Comparison between model RDCs (blue) and experiemt (green) of 10 EMW ensembles.

These ensembles were also examined for local conformational preferences at the K280 mutation site. This analysis differs from the previous study [49] in that it now incorporates the additional experimental data into the model and examines these conformational preferences within the context of K18. A single local conformational preference was identified for both WT

(Figure 27A) and ΔK280 (Figure 27B) forms of tau. Whereas there is a single local conformation preference in WT K18 corresponding to a bend centered at the mutation site, the corresponding region in the ΔK280 form exhibits a local conformational preference for a kinked, but more extended conformation. This is consistent with our previous hypothesis that a local conformational preference for more extended-like conformations in disease-associated states of tau may explain the differences in aggregation propensity and toxicity between the disease-associated mutant and WT forms of tau.

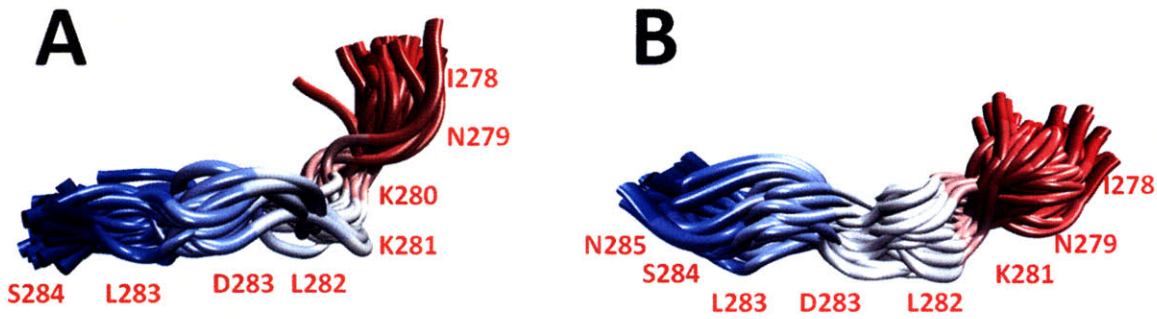


Figure 27 Local conformational preferences of segments at the K280 mutation site for (A) WT K18 and (B) ΔK280 K18.

Discussion

There are several possibilities as to why the segment model is unable to produce agreement with experiment. The underlying hypothesis of the model may be incorrect – there may exist subtle long-range correlations in the polypeptide chain that influence the distribution of conformations populated by the protein. If this is the case, local conformations cannot be sampled independently and the distribution of local conformational preferences may be altered by such long-range interactions. Another issue is that the current implementation uses a generic physical potential to obtain local conformational distributions, while in statistical coil models

(which have exhibited some success in qualitatively reproducing RDCs), the residue-dependent distributions are tuned to match RDCs from experiment. A better understanding of the physics of the unfolded state and potential functions specifically designed to model intrinsically disordered proteins may produce better agreement in the future. The implementation of the segment model may also require further refinement – further studies investigating the effect of segment size may be required. One may consider equilibrating the K18 structures using molecular dynamics to further relax the chain conformation after joining the segments. Finally, as discussed in the next chapter, there are limits to the accuracy of the current mapping methods between model structures and experiment which need to be addressed. These issues represent potential future avenues for these studies.

Further studies using EMW models of K18 should utilize additional structures to better fit the RDC data. As Figure 26 shows, disparities remain between experimental and model RDC signals. Furthermore, at the present, RDC measurements are only available for the wild-type form of K18. The availability of RDC measurements for disease-associated mutant forms of K18 in the future will improve the quality of comparison between wild-type and disease associated forms of K18. Nevertheless, these findings represent a first step in modeling the entire tau microtubule-binding repeat domain, and provide additional support for our previous hypothesis that the ΔK280 mutation affects the local conformational preference near the PHF6* aggregation-initiation site by favoring extended local conformations.

Methods

Sampling Conformations of K18 with the Segment Model

Replica Exchange Molecular Dynamics of Segment Peptides

The setup of the molecular dynamics simulations used to generate segment conformation distributions is comparable (and partially inspired by) to simulations by Ho and Dill in which replica exchange simulations were performed on peptide fragments of folded proteins [210].

A local sequence size of 8 residues was chosen for the size of the peptides used in the segment simulations, approximately the average persistence length of a polypeptide [67]. The sequence of K18 was divided into 26 peptides of 8 residues each, with an overlap of 3 residues between adjacent segments. By varying the length of the peptide, one effectively changes the length-scale of local interactions being accounted for, moving along the spectrum of model detail represented in Figure 24. Using a local sequence size of 2 residues, one obtains a statistical coil model where the phi/psi propensities are defined by the energy function. Using a local sequence size that is equal to the length of the protein, the simulation is effectively a replica exchange simulation of the entire protein. The ultimate choice of the segment size is a parameter of the model, and our choice was made to strike a balance between these two extremes. Furthermore, a similar replica exchange protocol has been used to sample conformations of 8 residue peptides in a previous study [210].

Each segment was simulated using 10ns of replica exchange molecular dynamics [73]. The first 5ns of REMD simulation was discarded as equilibration and only the last 5ns of simulation was used for the conformational distribution. Previous studies showed that backbone entropy of peptides of this size typically equilibrates within 3.5ns or less [210]. REMD simulations were

run in heat baths exponentially spaced between 260K and 700K. Exchanges were performed every 1ps. Inspection of the REMD trajectories confirmed that exchanges frequently occurred between all temperatures. Structures from the room temperature (298K) heat bath are saved prior to each exchange, generating 5000 structures for each segment. 26 segments are required to cover the entire sequence of K18, and thus 130,000 segment conformations are generated in total.

Constructing K18 Structures from Peptide Fragments

The simplifying approximation of the segment model is that conformations of sequentially-distal regions of the protein are independent. Thus, structures of K18 are constructed by independently sampling and joining peptide conformations of segments of the K18 sequence. This scheme is comparable to the structure-generation method in statistical coil models [68]. However, instead of building protein structures one residue at a time, the sequence is extended by independently sampling and adding one peptide-segment at a time. Starting with the N-terminal segment, each subsequent segment is independently sampled from its REMD trajectory and aligned using the backbone atoms of the 3 overlapping residues. For each K18 structure sampled using this method, a PDB file is created with duplicate atoms erased and residues renumbered.

Each K18 structure was minimized to remove bad contacts using 1000 steps of steepest descent minimization followed by 1000 steps of adopted basis newton-raphson minimization. Inspection of the resulting structures showed that this minimization protocol removes bad contacts while preserving the overall topology.

Computing Experimental Quantities

Backbone radius of gyration, chemical shifts, and residual dipolar couplings were computed for each K18 structure. Ensemble averages were compared with experimental measurements of chemical shifts [123, 124], residual dipolar couplings [19], and the radius of gyration [18].

Backbone radius of gyration was computed in CHARMM [157]. Residual dipolar couplings were computed by using REDUCE [211] to add non-polar hydrogens and PALES [82] to compute RDCs from the resulting all-atom structures. Chemical shifts were computed using SHIFTX (SHIFTX includes a built-in procedure to add missing non-polar hydrogen atoms) [2].

Computing ensemble averages across structures is performed as previously described [49]. In the segment model, each sampled K18 conformation is given equal weight. For ensemble averages of RDCs, a single scaling parameter is used to rescale the range of the RDC signal, as has been done in previous studies [67, 68].

Generation and Analysis of EMW Ensembles for K18

EMW was used to generate 70 ensembles of K18 consistent with experiment as previously described [49]. For WT K18, ensembles are fit against the experimentally measured RDC signal [19]. The resulting ensembles were consistent with measured Cα, CO, and H chemical shifts to within prediction error. Only the 30 ensembles closest to the experimentally measured RGYR were used for further analysis. Since RDC measurements are not available for ΔK280, chemical shift measurements were used for the optimization of ensembles. As with the WT ensembles, the 30 ΔK280 ensembles closest to the experimentally measured RGYR for K18 are used for further analysis.

Since K18 structures are generated from the local distributions, the identification of local conformational preferences was performed by clustering the peptide conformations using

MMTSB [212] with a 2.5Å cluster cutoff and identifying local conformational segment clusters present in K18 structures across all ensembles.

Chapter 6: Future Work

This work, in conjunction with recent experimental studies, represents a series of early approaches to characterizing structural details of tau [18, 19, 122-124]. The construction of more detailed models of tau protein will be facilitated by improvements to general methods for the characterization of IDPs.

The mapping between model structures and experimental measurements is a key step in the modeling of intrinsically disordered proteins such as tau. To determine ensemble averages from calculated ensembles one needs to first calculate properties from individual structures. Significant uncertainty associated with mapping individual structures to experimental properties is clearly undesirable as this leads to uncertainty in knowing how well the ensemble agrees with the experimentally determined result, which corresponds to an ensemble average. For example, most commonly used structural chemical shift prediction methods have an associated error on the order of 1ppm, but the error can vary widely depending on the atom type [2]. Measured chemical shifts of intrinsically disordered proteins often deviate from their random coil values by less than 1ppm and these deviations are often interpreted as indicative of residual conformational preferences [123]. Thus, at present the mapping between model structures and experimental values introduces a large degree of uncertainty with regards to model optimization, interpretation and validation.

Another difficulty in modeling an unfolded ensemble is the degeneracy of the solution space. Given a limited amount of structural information there exist multiple structural ensembles consistent with the available experimental data. Since inclusion of additional independent measurements decreases the size of the solution space, one approach to this problem is to obtain

a large number of independent measurements (although the effect of different chemical environments on the structural ensemble can complicate data interpretation). However, what is even more problematic is that given the large size of the solution space, it is likely that degeneracy will be a problem for IDPs even when a relatively large set of experimental data are used.

Physics-based models can play a role in addressing the problem of solution degeneracy. By including constraints imposed by knowledge of the underlying physics of the system, one can both reduce the space of solutions as well as quantitatively examine whether those physical assumptions are consistent with the experimental data, thus gaining deeper insight into the properties of the ensemble. The segment model is an attempt to account for physical interactions determining local conformational preferences while still allowing for efficient generation of conformers. While additional work is necessary to enable such models to predict experimental quantities, I believe that such approaches will enable greater insights into the unfolded state of intrinsically disordered proteins and potential mechanisms of aggregation.

Finally, the most important extension of this work is to utilize these structural insights in the design of therapeutics which target tau. We have proposed potential approaches for doing so in chapter 2. Other recent reviews have also discussed strategies for therapeutic targeting of intrinsically disordered proteins [129]. Perhaps the most critical endpoint for these structural characterizations is to establish a basis for the rational design of therapeutics for neurodegenerative diseases.

Appendix: Residual structure within the disordered C-terminal segment of p21Waf1/Cip1/Sdi1 and its implications for molecular recognition

(Note: Experimental Sections of this Appendix were performed and written by collaborators Mi-Kyung Yoon, Byong-Seok Choi, and James J. Chou. Other sections of this were written collaboratively with Veena Venkatachalam, an undergraduate student. This work was published as M.-K. Yoon, V. Venkatachalam, A. Huang, B.-S. Choi, C. Stultz, and J. Chou, "Residual structure within the disordered C-terminal segment of p21Waf1/Cip1/Sdi1 and its implications for molecular recognition," Protein Science., vol. 18, pp. 337-347, 2009.)

Abstract

Probably the most unusual class of proteins in nature is the intrinsically unstructured proteins (IUPs), because they are not structured yet play essential roles in protein-protein signaling. Many IUPs can bind different proteins, and in many cases, adopt different bound conformations. The p21 protein is a small IUP (164 residues) that is ubiquitous in cellular signaling, e.g., cell cycle control, apoptosis, transcription, differentiation, and so forth; it binds to approximately 25 targets. How does this small, unstructured protein recognize each of these targets with high affinity? Here, we characterize residual structural elements of the C-terminal segment of p21 encompassing residues 145 – 164 using a combination of NMR measurements and molecular dynamics simulations. The N-terminal half of the peptide has a significant helical propensity which is recognized by calmodulin while the C-terminal half of the peptide prefers extended conformations that facilitate binding to the proliferating cell nuclear antigen (PCNA). Our results

suggest that the final bound conformations of p21(145-164) pre-exist in the free peptide even without its binding partners. While the conformational flexibility of the p21 peptide is essential for adapting to diverse binding environments, the intrinsic structural preferences of the free peptide enables promiscuous yet high affinity binding to a diverse array of molecular targets.

Introduction

Many proteins adopt a well-defined tertiary structure under physiological conditions, and this structure largely determines protein function. However, there is a class of proteins known as intrinsically unstructured proteins (IUP) that also plays specific roles in protein-protein recognition. Some well-known examples include the phosphorylated kinase-inducible domain (pKID) of the cAMP responsive element binding protein (CREB) [213], the transcriptional activation domain (TAD) of p53 [214], and the GTPase-binding domain (GBD) of the Wiskott–Aldrich syndrome protein (WASP)[215]. Spectroscopic studies suggest that these intrinsically unstructured proteins are not completely random, but can exhibit residual secondary structural preferences. For example, NMR studies demonstrated that the linker helix of $p27^{Kip1}$ has a nascent secondary structure in its free state [216] although it is largely unstructured in solution [217]. It is increasingly apparent that residual structure of intrinsically unstructured proteins (IUPs) plays crucial roles in molecular recognition [22, 218, 219]. In this regard, it has been suggested that the classic protein structure-function paradigm for IUPs be re-assessed and that protein function for these systems can be understood using a formalism that models the IUP structure as an ensemble of distinct conformations [22, 220, 221].

$p21^{Waf1/Cip1/Sdi1}$ (hereafter referred to as p21) is an IUP involved in the regulation of the cell cycle [222]. p21 was first identified as a cyclin-dependent kinase (Cdk) inhibitor [223] that

mediates the G1/S arrest [224] and later was found to function in apoptosis [225], differentiation [226], transcription [227], DNA synthesis control [228] and stem cell self-renewal [229]. The C-terminal region of p21, which is unique among the Cip/Kip family of Cdk inhibitors, interacts with a large array of proteins, including the proliferating cell nuclear antigen (PCNA) [230, 231], calmodulin (CaM) [232], SET [233], c-Myc [234], and E7 oncoprotein of human papilloma virus 16 (HPV-16) [228].

How does p21, a small protein of 164 residues, physically recognize so many structurally dissimilar proteins without sacrificing binding affinity? The binding diversity of the C-terminal segment of p21 has been attributed to its ability to acquire different conformations upon binding to distinct targets [235]. Although the far UV CD spectra suggest that residues 145-164 (p21(145-164)) is unstructured, this peptide adopts a well defined structure having a helical N-terminal region and an extended strand C-terminal region when bound to PCNA [15]. Based on previous examples of CaM-substrate binding [236] and CD measurement of mutant p21(145-164) [235], p21(145-164) is likely to acquire a helical conformation when bound to CaM.

In this study, we combine NMR measurements of free p21(145-164) with molecular dynamics (MD) simulations to obtain models for the unfolded ensemble of the free peptide at a physiologically relevant temperature and pH. We found that the N-terminal half of the peptide has a significant amount of residual helical structure and the C-terminal half has a preference for extended conformations in the unbound state of p21(145-164). NMR dipolar coupling measurements of the CaM – p21(145-164) complex indicate that the peptide is helical when bound to CaM, which in turn suggest that the region of peptide with helical preference is likely to interact with CaM. On the other hand, the C-terminal loop-like region of the peptide adopts an extended conformation when bound to the PCNA [15]. Our results show that the structure

adopted by p21(145-164) upon binding to CaM or PCNA already exist in the free peptide in significant population and suggest that the pre-formed structural elements of p21(145-164) contribute to its binding specificity.

Results

Residual secondary structure in p21(145-164) detected by NMR spectroscopy.

The ^{1}H-^{15}N HSQC spectrum of ^{15}N-labeled p21(145-164) (Figure 28) showed poor dispersion (< 1 ppm) of amide proton chemical shifts, consistent with a largely unstructured peptide as observed by far-UV CD spectroscopy [235]. To investigate whether there is any residual structure in p21(145-164), we measured $^{13}C^{\alpha}$ and $^{1}H^{\alpha}$ chemical shifts, as well as the 3-bond $^{3}J_{HN\text{-}H\alpha}$ coupling constants. The $^{13}C^{\alpha}$ and $^{1}H^{\alpha}$ chemical shift values are very sensitive to local conformation and thus their deviations from the values of random coil, known as secondary chemical shifts, are indicative of secondary structure [237]. Secondary shift analysis is the most widely-used method for detecting residual structural elements in largely unstructured polypeptide chains [238-240]. The $^{13}C^{\alpha}$ secondary chemical shift of p21(145-164) (Figure 29A) shows that the N-terminal half of the peptide, encompassing residues Met147 – Lys154 is partially helical on average, while its C-terminal half shows a preference for extended conformations. In agreement with the $^{13}C^{\alpha}$ shifts, the $^{1}H^{\alpha}$ secondary shifts (Figure 29B) are consistent with an increased propensity for helical conformation for the N-terminal segment Thr148 – Arg155. However, the $^{1}H^{\alpha}$ shifts for the C-terminal half of the peptide are not characteristic enough to draw any conclusions on the secondary structure preferences. Independent from the chemical

shifts, the deviations of the $^{3}J_{HN\text{-}H\alpha}$ coupling constants from random coil values (Figure 29C) also suggest helical tendency of the N-terminal half of the peptide, Thr145 – Lys154.

The experimental chemical shifts and scalar coupling constants together indicate that the N-terminal half of the peptide is helical, on average, under physiological conditions. While the $^{13}C^{\alpha}$ chemical shifts of residues Leu157 – Lys163 suggests that this region is extended on average (Figure 29A), the $^{1}H^{\alpha}$ secondary shifts (Figure 29B) and the $^{3}J_{HN\text{-}H\alpha}$ deviations (Figure 29C) are less conclusive.

Modeling the unfolded state of p21(145-164) with MD simulations.

NMR measurements for unfolded proteins correspond to ensemble averages over a number of structurally dissimilar states and therefore do not provide information about the underlying distribution of conformers in the ensemble. To further our understanding of the conformers that make the unfolded ensemble, we used molecular dynamics (MD) simulations to generate structural ensembles which agree with our experimental data.

Our method, energy-minima mapping and weighting (EMW), associates a statistical weight with each structure that corresponds to the probability that the given protein samples that conformation. The application of EMW presented here is similar to that previously described [49] and relies on obtaining NMR chemical shift data for a free peptide in solution, and optimizing structural ensembles containing energetically favorable conformations of that peptide to minimize the error between the calculated chemical shifts and the experimentally measured values. The resulting ensembles agree with the experimental values, while also fulfilling physical constraints imposed by the potential energy function.

As the construction of an unfolded ensemble is an underdetermined problem, there may be several ensembles that agree with any given set of experimental constraints. Therefore EMW does not strive to construct one ensemble that models the unfolded state of p21(145-164). Instead, we generate multiple ensembles that are all consistent with a given set of experimental data, and focus our analysis on local structural motifs that are present across ensembles.

EMW was used to construct 250 ensembles using absolute $^{13}C^{\alpha}$ chemical shifts. The 250 ensembles were then ranked according to their ability to reproduce experimentally determined amide nitrogen chemical shifts, which were not used in the optimization procedure. The ten ensembles best able to reproduce amide nitrogen chemical shifts were chosen for further analysis. Calculated $^{13}C^{\alpha}$ chemical shifts for these ten ensembles were in excellent agreement with experiment (Figure 29), and nitrogen chemical shifts were all within 1.5ppm of the experimental values (an error comparable to that of available chemical shift prediction algorithms [2]). $^{1}H\alpha$ chemical shifts were also computed for these ten ensembles, which resulted in errors within 0.2ppm, which is also comparable to the error associated with available chemical shift prediction algorithms [2].

To identify local conformations preserved across ensembles, all structures were clustered based on the rms backbone deviation of contiguous 6-residue subsequences in p21(145-164). A characteristic length of six residues was chosen as the local region size for this analysis since a crystal structure of a bound state of p21(145-164) contained local structured regions approximately six residues in length [15]. Clustering all contiguous 6 residue segments resulted in 225 distinct clusters. Each cluster is representative of a local conformation within p21(145-164). The total weight associated with a given cluster in an ensemble is given by the sum of the weights of all structures in the cluster. A given cluster is said to be preserved across all ten

ensembles if it has a non-zero weight in each ensemble. Using this definition, only 5.8% of the clusters were preserved across all ensembles.

A preserved structural motif that is present in all independent ensembles is likely required to reproduce the experimental data. Consequently, we consider such locally preserved structures to represent local conformational preferences. Structures of conserved local conformations offer a more detailed view of ensemble characteristics than ensemble averaged experimental secondary chemical shifts.

Conformational preferences for p21(145-164) are shown in Figure 31. Several points are clear from Figure 31. First, every residue in p21(145-164) is found in a 6-residue segment that has an extended conformation. Hence, the simulations predict that each residue can adopt an extended state in solution, including the N-terminal residues that have positive secondary Ca chemical shifts (Figure 29A). What distinguishes the N-terminus is the fact that these residues can also adopt helical conformations (Figure 31). Three conserved N-terminal helical clusters were found, corresponding to preformed helical states in the unfolded ensemble, in the six-residue regions corresponding to residues 147-152, 148-153, and 149-154. Residues 153-158 can also adopt a loop/turn conformation in solution. Lastly, the simulations predict that residues 159-163 have a distinct preference for only extended states.

It was previously reported that residues (146-151) in PCNA-bound form of p21(139-160) forms a 3_{10} helix, while the C-terminal region spanning residues 152-160 adopts an extended strand which hydrogen bonds to a neighboring β strand in PCNA (Figure 32A, B) [15]. We sought to determine whether comparable local conformational preferences are predicted for the unfolded ensemble representing the unbound state. We find that the 6 residue subsequence ranging from residues 147 to 152 can adopt a helical conformation in solution and that the

subsequence consisting of residues 155-160 adopts an extended state, suggesting that there local structural preferences in the unfolded ensemble similar to those adopted by the bound form of p21(139-160) (Figure 32C).

Helical mode of p21(145-164) binding to Ca^{2+}-calmodulin from NMR dipolar couplings.

Based on CD measurements of mutant p21(145-164) [235] and the previous knowledge of CaM-binding peptides [236], we expected p21(145-164) to be helical upon binding to CaM. Previously, at least 180 CaM recruitment signaling (CRS) motifs were identified and classified based on the spacing of the hydrophobic residues of CRS that make major hydrophobic interaction with CaM [236]. However, p21 peptide contains no sequence that conforms to any of these known CRS motifs.

In order to confirm the helical mode of p21(145-164) binding to Ca^{2+}-CaM, two types of residual dipolar couplings (RDCs), $^1D_{NH}$ and $^1D_{C\alpha H\alpha}$, were measured for Ca^{2+}-CaM in complex with p21(145-164) in a liquid crystalline medium containing 18 mg/ml filamentous phage Pf1. RDCs have been successfully used to determine the binding mode of CaM-interacting peptide to Ca^{2+}-CaM [241, 242]. The experimentally measured RDCs of CaM in complex with p21(145-164) were fitted to the free Ca^{2+}-CaM structure (Figure 33A) and CaM/peptide complexes using the singular value decomposition (SVD) method [243]. The measured RDCs show very poor agreement with the dumbbell-shape crystal structure of the free Ca^{2+}-CaM (Figure 33A) because the conformation of Ca^{2+}-CaM bound to p21(145-164) is very different from that of free Ca^{2+}-CaM. Among the 13 CaM/peptide complexes tested, including the complexes that show different types of binding mode, the best correlation was obtained for CaM/CaMKII complex that belongs

to the 1-10 class [244] (Figure 33B). This suggests that p21(145-164) adopts a known CaM-binding mode (possibly the 1-10 class) in which the CaM-interacting region is helical.

Discussion

The NMR data and molecular dynamics simulations are used to explain the binding promiscuity and specificity of p21(145-164) mediated protein-protein signaling. The binding promiscuity can be attributed to the structural plasticity, or unstructured nature, of the peptide, which allows it to adapt to the distinct structural environments of many different target proteins. A fundamental question remains - if the peptide is largely unstructured, how does it bind multiple proteins with high affinity?

We first examined the residual structure of p21(145-164) in solution and compared locally preferred conformations to regions of structure in bound forms of p21(145-164). The crystal structure of p21(139-160) bound to PCNA shows that residues Ser146 – Tyr151 of p21 forms a 3_{10}-helix which is involved in hydrophobic interaction with PCNA, whereas residues His152 – Ser160 of p21 adopts an extended β-strand conformation which interacts with PCNA [15]. We have shown the helical preference of the N-terminal region of the p21(145-154) which corresponds to the helical region of the PCNA-bound structure . The chemical shifts and scalar couplings for residues Arg155 – Lys163 show no sign of residual helical structure and the MD simulations suggest that the helical and extended portions of p21 that bind PCNA exist in the free state prior to binding (Figure 32). The recognition for PCNA in this case comes from the fact that residues His152 – Ser160 of p21 have a preference for conformational arrangements which readily expose the positively charged residues for specific interactions with PCNA. In the

extended, uncoiled conformation of the p21 peptide, residues Arg155 and Arg156 readily interact with the negatively charged Asp122 and Asp29 of PCNA, respectively.

In addition, the NMR structure of Cdk4-bound p21(141-160) also includes a helical region (residues 149-156) [16]. Similarly, model ensembles of the unbound state generated by EMW suggest that the region corresponding to residues 149-154 can adopt helical conformations (see Figure 31).

We then examined the mechanism of recognition between p21 and Ca^{2+}-CaM. Although the NMR resonances of the p21(145-164) peptide bound to CaM are extremely broad due to chemical exchange, the resonances of CaM are much less affected by the peptide and thus allow accurate structure measurement. An extensive set of RDCs measured for CaM bound to the p21(145-164) peptide indicates a CaM-substrate interaction mode in which the substrate adopts an α-helical conformation. For the free p21(145-164) peptide, the NMR chemical shifts and scalar couplings together show that residues 147 – 154 have significant helical propensity, and MD simulations suggest that the N-terminal region has a strong preference for helical conformations while helical structure is absent in the C-terminal region. We believe the residual helical segments observed in the free p21(145-164) peptide is responsible for its initial recognition with Ca^{2+}-CaM. The helical structure is then stabilized by the binding. Overall, the structural propensities of the free p21(145-164) peptide correlate well with the different structures adopted by p21 upon binding to different targets. This observation suggests that pre-existing residual conformations of p21 provide the initial recognition for the target proteins. Bindings then further stabilize these residual conformations. The p21 peptide provides an interesting example of how residual structural elements of an IUP are involved in specific and diverse protein-protein signaling.

Materials and Methods

Cloning, Protein Expression and Purification

The peptide p21(145-164) was expressed as a C-terminal in-frame fusion to the trpLE protein containing the N-terminal 9-His tag. A pair of Asp-Pro residues was engineered between trpLE and p21(145-164) for acid catalyzed cleavage to release the p21 peptide from the fusion protein. The expression vector was constructed by inserting the DNA fragment of p21(145-164) into the C264 vector, a gift from Dr. M.E. Call, Harvard Medical School, Boston.

Escherichia coli strain BL21 (DE3) that express the trpLE-fused p21(145-164) were cultured in M9 minimal media for isotope labeling. The cell cultures were grown at 37 °C to OD_{600} of 0.6-0.8 before overnight induction at 25 °C with 0.4 mM IPTG. Inclusion bodies were dissolved with a buffer containing 50 mM Tris, pH 8.0, 0.2 M NaCl, 6 M guanidine HCl, 10 mM imidazole. The cleared solution was bound to Ni^{2+} affinity column (Sigma) and eluted in 50 mM Tris, pH 8.0, 0.2 M NaCl, 6 M guanidine HCl, 400 mM imidazole. The eluted fusion protein was dialyzed against water to remove guanidine HCl. The precipitant was pelleted by centrifugation at 3000 rpm for 30 min. Incubation of the pellet in 0.1N HCl at 37 °C for 3 days released the p21(145-164) peptide from the trpLE fusion partner. The released peptide was dialyzed against water, lyophilized, and purified by reverse-phase HPLC on a C18 column (Grace-Vydac) with a gradient of water containing 0.1% trifluoroacetic acid (TFA) to acetonitrile containing 0.1% TFA. The resulting peptide was lyophilized and dissolved in 100 mM KCl, 10 mM $CaCl_2$, pH 6.5.

CaM was expressed and purified as previously described [245]. Isotropic NMR samples were prepared in 100 mM KCl, 10 mM $CaCl_2$, pH 6.5 in 93 % H_2O/7% D_2O. The aligned sample

contained 18mg/ml filamentous phage Pf1 (Asla Labs, Riga, Latvia), 100 mM KCl, 10mM $CaCl_2$, and 1mM sodium azide, pH 6.5 in 93 % H_2O/7% D_2O.

NMR Spectroscopy

All NMR spectra were collected at 30°C on Bruker and Varian spectrometers operating at ^1H frequencies of 500 MHz or 600 MHz and equipped with cryogenic probes. The sequence-specific backbone assignments were accomplished using pairs of HNCACB/CBCA(CO)NH and HNCA/HNCACB on the ^{15}N, ^{13}C-labeled CaM in complex with unlabeled p21(145-164) and ^{15}N, ^{13}C-labeled p21(145-164), respectively. Two types of backbone RDCs, $^1D_{NH}$ and $^1D_{C\alpha H\alpha}$, were measured on the ^{15}N, ^{13}C-labeled CaM in complex with unlabeled p21(145-164). The ^1H-^{15}N RDCs were obtained from $^1J_{NH}/2$ and $(^1J_{NH}+^1D_{NH})/2$, which were measured at 600 MHz (^1H frequency) by interleaving a regular gradient enhanced HSQC and a gradient-selected TROSY, both acquired with 80 ms of ^{15}N evolution. The ^1H-$^{13}C^{\alpha}$ RDCs were measured at 600 MHz (^1H frequency) using the 3D CT-(H)CA(CO)NH without ^1H-decoupling [246]. Measurement of $^3J_{HN\text{-}H\alpha}$ coupling constant for determining backbone ϕ angle was carried out on the ^{15}N, ^{13}C-labeled p21(145-164) using the 3D HNHA experiment [247]. The ^1H chemical shifts were referenced directly to external 2,2-dimethyl-2-silapentane-5-sulfonic acid (DSS) in D_2O and ^{13}C chemical shifts are indirectly referenced to 0 ppm proton using the method in Wishart et al. [248].

Data processing and spectra analyses were done in NMRPipe [249], CARA [250], and Sparky (http://www.cgl.ucsf.edu/home/sparky). RDCs were extracted by subtracting isotropic couplings from the aligned couplings. Fitting of RDCs to structures was done by singular value decomposition [243], using the program PALES [82]. The goodness of fit was assessed by both

Pearson correlation coefficient (R) and the quality factor (Q) {}. The ^{1}H-$^{13}C^{\alpha}$ RDCs were normalized to the ^{1}H-^{15}N RDCs by a scaling factor of 0.5.

Molecular dynamics simulation

Energy-minima Mapping and Weighting (EMW)

To construct an ensemble that represents the unfolded state of p21(145-164), we employ EMW method [49]. Details of EMW are described in detail below.

(i) Conformational Sampling

The goal of conformational sampling was to generate a library of energy-minimized structures with representatives from all regions of conformational space accessible to the peptide. This was done using quenched molecular dynamics (QMD) (6). To ensure that both compact and extended structures were adequately sampled, QMD was carried out at 50 different end-to-end distance constraints, spanning a range from 4Å to 53Å. At each distance constraint, a polar hydrogen model of an extended peptide having the sequence TSMTDFYHSKRRLIFSKRKP (p21(145-164)) was constructed using CHARMM, and a harmonic penalty was introduced to enforce the desired distance between $^{13}C^{\alpha}$ of T145 and $^{13}C^{\alpha}$ of P164. The structure was then minimized using 500 steps of steepest descent minimization followed by 10000 steps of minimization using the Adopted Basis Newton Rhapson algorithm. Next, the structure was heated to 1000 K for 10 ps and allowed to equilibrate for 10 ps, before high temperature MD was run for 3 ns. Throughout the simulation, a Berendsen heat bath was used to maintain the temperature [159], and the EEF1 energy function (a Gaussian solvent exclusion model for the solvation free energy) was used to assign energies [140]. The SHAKE algorithm was employed to hold bonds to hydrogen atoms fixed near their equilibrium values, allowing for a 2 fs time step

during high temperature MD simulations [160]. The peptide's coordinates were saved after every 5000 steps (10 ps) of high temperature MD simulation, yielding 300 structures per end-to-end distance constraint. Thus, 15000 structures were created using high-temperature molecular dynamics.

With end-to-end distance constraints still in place, each of these structures was coupled to a Berendsen heat bath at 298 K and cooled for 40 ps, at which point the end-to-end constraint was removed and the system was minimized using 10000 steps of Adopted Basis Newton Rhapson minimization [159]. Cooling and equilibrating each structure before minimization gave the system a chance to escape shallow local energy minima, thereby making more stable structures accessible. The 15000 structures obtained in this manner comprised our structure library.

(ii) Model Optimization

A goal of this procedure is to find ensembles that represent the solubilized p21 peptide, where each ensemble consists of 15 structures and their associated weights. Accordingly, the optimization stage of EMW involved generating such ensembles by choosing structures from the conformational library generated in the first step and assigning weights to these structures. Experimentally determined $^{13}C^{\alpha}$ NMR chemical shifts for the peptide were used to determine what constituted an optimal ensemble; structures and weights were assigned to minimize the root mean square error between $^{13}C^{\alpha}$ chemical shifts computed from the model using SHIFTX and $^{13}C^{\alpha}$ chemical shifts that were experimentally measured [2]. This was done by minimizing an appropriate error function, f, given by:

$$f\left(\{\omega_i, X_i\}_{i=1}^{N}\right) = \sum_{j=1}^{r}\left(\sum_{i}^{\mathrm{N}}\left(\omega_i S_{c_\alpha}^{X_i}(j)\right) - S_{c_\alpha}^{Exp}(j)\right)^2 \qquad \text{(A.1)}$$

where N is the number of structures in the ensemble (N = 15), X_i is the i^{th} structure, ω_i is the weight of the i^{th} structure, r is the number of residues in the peptide for which experimental chemical shift data is available (r = 18), $S_{c_\alpha}^{X_i}(j)$ is the calculated $C\alpha$ chemical shift of residue j in structure X_i, and $S_{c_\alpha}^{Exp}(j)$ is the experimentally determined $^{13}C^{\alpha}$ chemical shift of residue j [49].

A simulated annealing protocol using a cooling schedule based upon that described by Nulton et al. was implemented [207]. Each ensemble was generated from an initial ensemble consisting of 15 Boltzmann-weighted structures chosen at random from the conformational library [48]. This initial ensemble was subjected to an iterative simulated annealing protocol that minimized the rmse between measured and predicted $^{13}C^{\alpha}$ chemical shifts. Each step of the annealing protocol consisted of carrying out Monte Carlo steps at a given value of the control parameter T, which is analogous to the temperature in physical systems, until the system had equilibrated.

A Monte Carlo step consists of perturbing the ensemble by replacing one conformer in that ensemble with a conformer from our structure library. Weights for all structures are then reassigned to minimize the overall error, f. The number of Monte Carlo steps for a given value of the control parameter, as well as the schedule used to decrease the overall temperature, is as described in a previous work [206]. Overall, 250 ensembles were generated using this simulated annealing protocol.

(iii) Model Validation

The rmse between predicted nitrogen chemical shifts and measured nitrogen chemical shifts for each ensemble was calculated for each of the 250 models generated, and those ensembles in which this error was less than 1.5 ppm were taken to be valid models based upon their ability to predict experimentally measured amide nitrogen chemical shifts. Ten valid ensembles were

found that reproduced the NMR chemical shift data well, due to the underdetermined nature of the problem, but we accounted for this by using all ten of these independently generated, validated structural ensembles in our analysis and focusing on those structural motifs that were preserved across all of them.

Procedure for Identifying conserved preformed structures

Two different methods were employed to find structural motifs that were present across all validated model ensembles of the unfolded state. Since studies have shown that bound states of p21(145-164) adopt helical conformations, we looked for evidence of preformed helical motifs in the unfolded ensemble. We then sought to identify other structural motifs suggested by the model ensembles.

Identification of conserved local structure by clustering

We sought to identify other types of local structural motifs in the peptide. To this end, all conformers were clustered based on local conformational preferences. Since the helical region in the crystal structure of PCNA-bound p21 is six residues in length, we defined the characteristic size of a local structural motif to be six residues. To account for all local motifs, every six residue subsequence of p21(145-164) was analyzed to find preserved conformations. This was accomplished by clustering based upon backbone atom RMS deviations within each six-residue window of interest. Clustering was carried out in MATLAB (© Mathworks) such that the maximum rmsd between any two structures was 2.5 Å. Clustering based on other window sizes (5, 7, and 8 residues in length) was performed to ensure that analysis was relatively insensitive to the choice of window size. Clusters that were represented in all ten ensembles were identified as preserved local structural motifs.

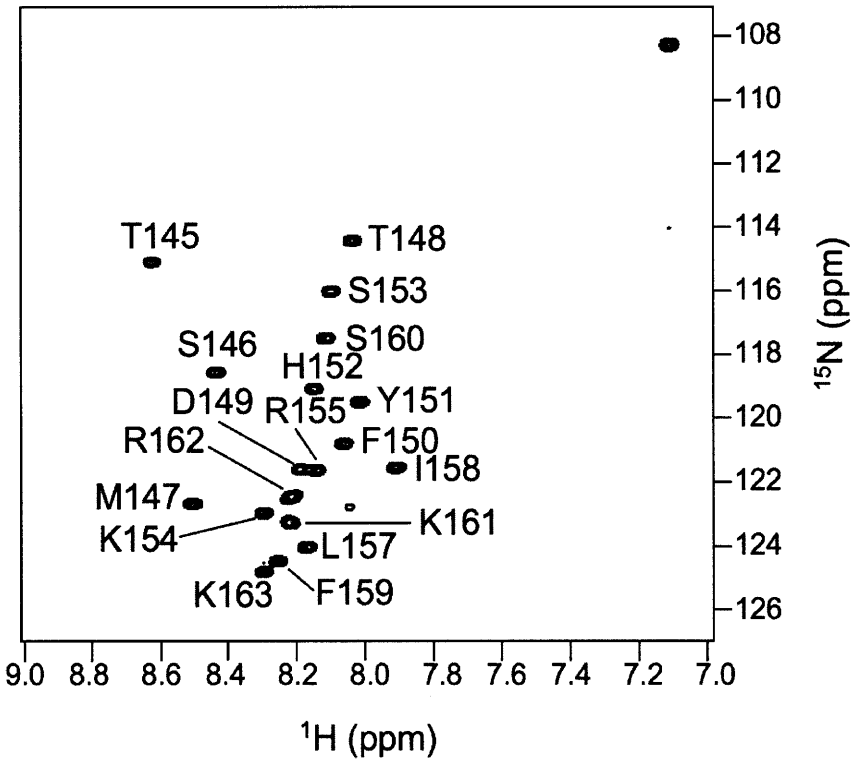


Figure 28 The ^{1}H-^{15}N heteronuclear single quantum correlation (HSQC) spectrum of the ^{15}N-labeled p21(145-164) recorded at ^{1}H frequency of 500 MHz, pH 6.5, and 30 °C. All of the backbone $^{1}H^{N}$ and ^{15}N are assigned except for Arg156.

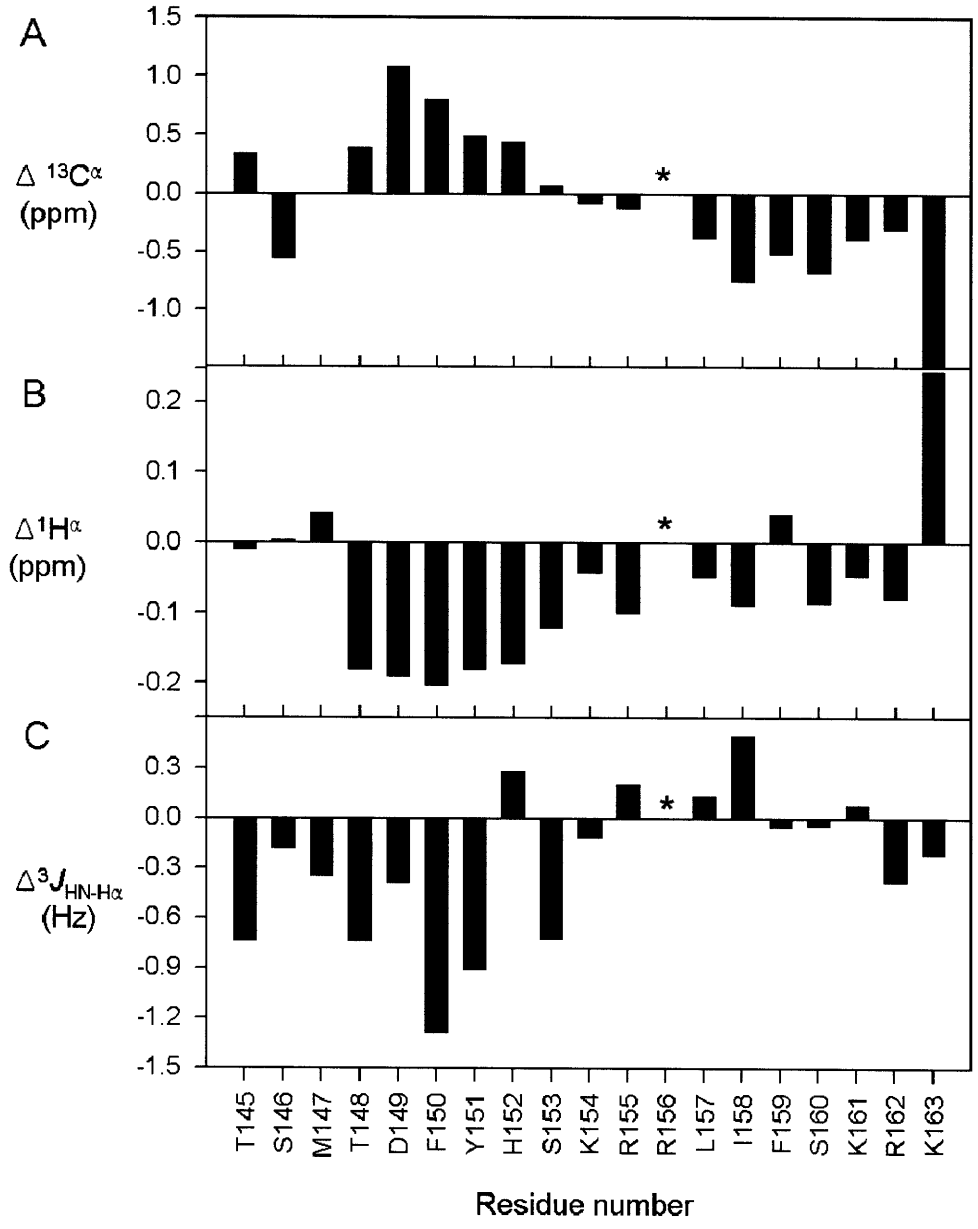


Figure 29 NMR measurements of residual secondary structures of p21(145-164). (A) Deviation of $^{13}C^{\alpha}$ chemical shifts of p21(145-164) from the random coil values [251]. $\Delta^{13}C^{\alpha}$ = $^{13}C^{\alpha}$ (p21) - $^{13}C^{\alpha}$ (random coil). (B) Same as in (A) for the $^{1}H^{\alpha}$ chemical shifts. (C) Deviation of the $^{3}J_{HN\text{-}H\alpha}$ coupling constants of p21 (145-164) from random coil values [252]. $\Delta^{3}J_{HN\text{-}H\alpha} = {}^{3}J_{HN\text{-}H\alpha}$ (p21) - $^{3}J_{HN\text{-}H\alpha}$ (random coil). Data are not available for Arg156 (indicated by *).

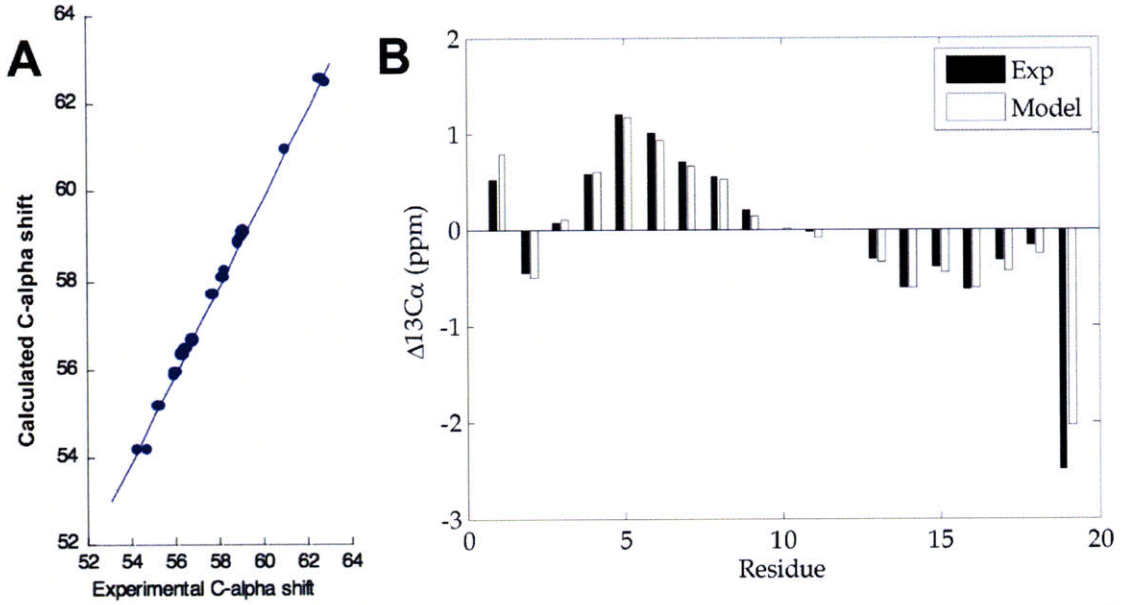


Figure 30 Comparison of experimental Cα chemical shifts and Cα chemical shifts calculated using the "worst" model; that is, the model with the highest rms difference between the calculated and experimental result. (A) Relationship between experimental and calculated absolute Cα chemical shifts. (B) Comparison of experimental and calculated Cα secondary chemical shifts.

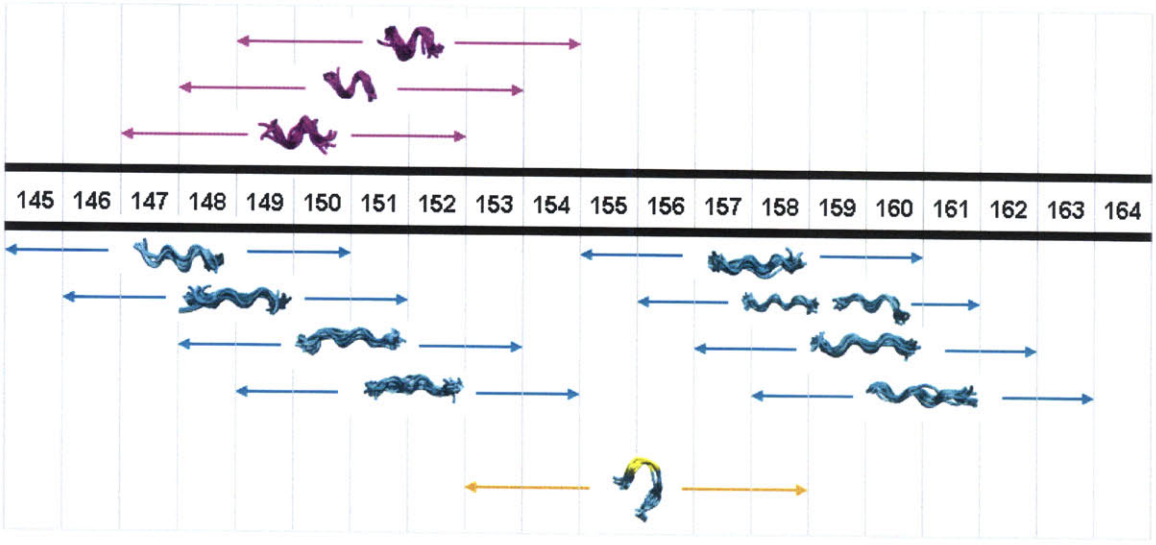


Figure 31 Conserved 6-residue structural motifs. Helical conformations are colored purple and extended regions are shown in cyan. The region corresponding to a turn is colored yellow.

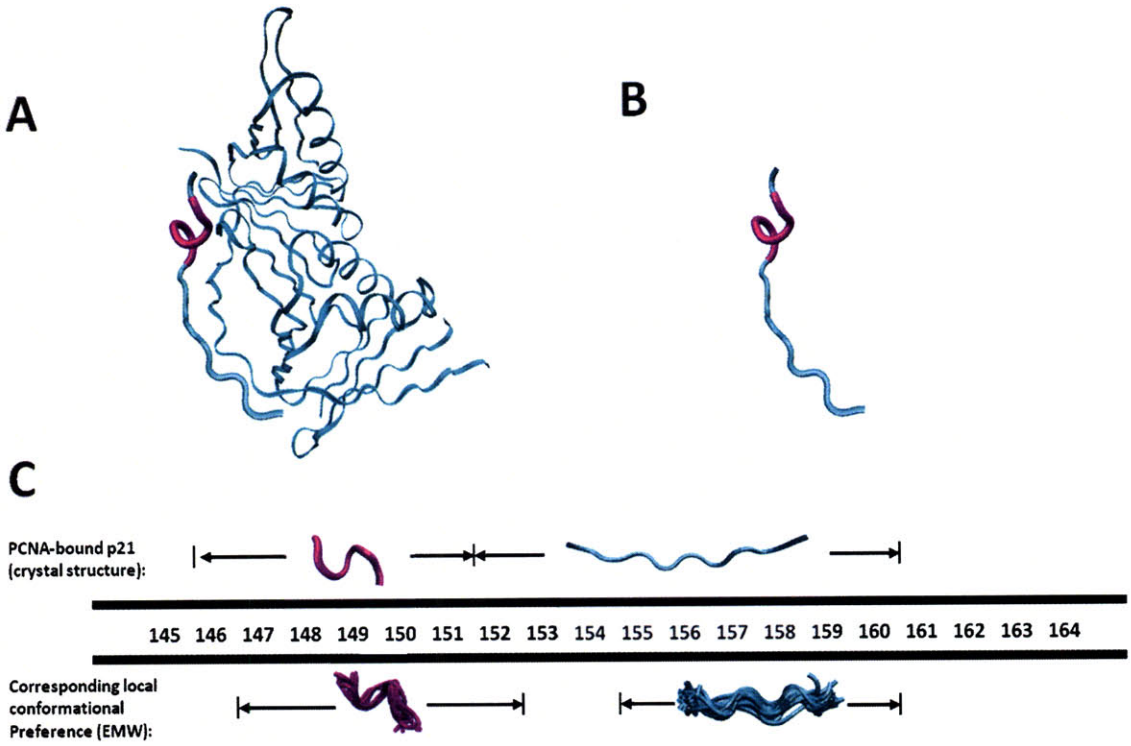


Figure 32 A) X-ray crystallographic structure of a p21 model peptide (residues 139-160) bound to PCNA [PDB ID 1AXC [253]]. The helical region (p21 residues 146 – 151) is highlighted in purple, while the extended C-terminus (p21 residues 152 – 160) is depicted in cyan. Only residues of p21 included in our model peptide (145-160) are shown. B) Structure of the p21 model peptide alone. C) Comparison of corresponding local conformational preferences in models of the unbound form of p21(145-164) . Regions for which models of the unbound form include local conformations which match the bound form crystal structure are indicated.

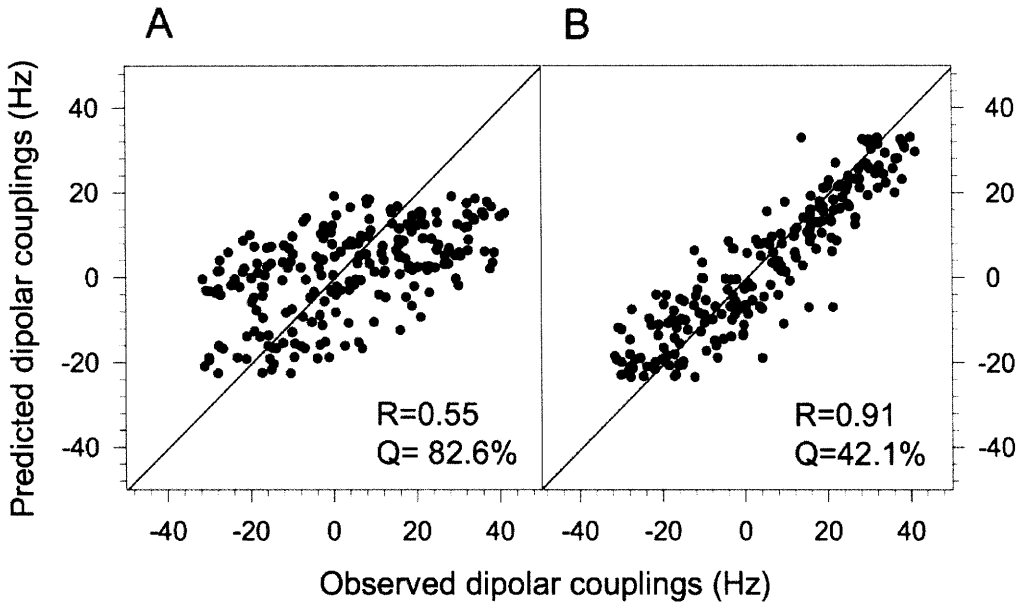


Figure 33 Determination of the CaM/p21 binding mode using RDCs. Best-fitting of observed backbone RDCs of the Ca^{2+}-CaM/p21(145-164) complex to (A) the free Ca^{2+}-CaM crystal structure (PDB code: 1EXR) and (B) the crystal structure of Ca^{2+}-CaM/CaMKII (PDB code: 1CDM). All RDCs are normalized to $^1D_{NH}$.

Acknowledgements

> "Why is it precisely at this *intermediate level* in the hierarchy of successively superimposed unities (cell, organ, human body, state) – why, I ask, is it precisely at the level of my body that unitary self-consciousness comes into the picture, whereas the cell and the organ do not as yet possess it and the state possesses it no longer? Or, if this is not so, how is my self constituted out of the individual selves of my brain cells? Is there a higher Self similarly constituted out of the consciousness of myself and my fellow-men, equally and directly conscious of itself as a unity – the Self of the state or of the whole of humanity? ... Such questions vanish as soon as the root of that directly experienced unity which leads to the hypothesis of the Self is transferred to the metaphysical unity, the essential uniqueness of consciousness in general. The categories of *number*, of *whole* and of *parts* are then simply not applicable to it; the most adequate, though no doubt still somewhat mystical expression of the situation being this: the self-consciousness of the individual members are numerically identical both with each other and with that Self which they may be said to form at a higher level; each member is in a certain sense justified in saying 'L'État, c'est moi'."
>
> -Erwin Schrödinger, My View of the World

I would not have had the opportunity to pursue a graduate degree if not for the support of my family. To the elder generation in my family – my mother, father, aunts, uncles, and grandparents: my respect for you grows every day. Truth be told, the accomplishment of obtaining a doctorate is dwarfed by the accomplishment of forging a new life in a foreign country. This work is in many respects only the most recent consequence of your lifetimes of labor and love.

I am very appreciative of the guidance I have received from my research advisor, Collin M. Stultz. I never have any doubt that he places the best interests of his students above all else. Through countless instances of feedback in the contexts of discussions, presentations, writing, and spirited debate, he deserves much of the credit for my maturation as a scientist.

I would also like to thank Bruce Tidor and Amy Keating for their patient guidance as members of my thesis committee. I have also been fortunate enough to encounter other intelligent and selfless mentors throughout my academic and professional career – Shagi-Di Shih, Simik Sarkis-Kelly, Kimmen Sjolander, Anne Nesbet, Dennis Freeman, Tomas Lozano-Perez, Paul T. Matsudaira, and Eric Grimson. I hope to continue to having enlightening discussions with all of these mentors as I move on to the next phase of my professional life.

This work was immensely aided by a series of experimental results published by Marco D. Mukrasch, Daniela Fischer and the group of Markus Zweckstetter, who have been extremely generous in sharing the raw data from their publications.

To all my friends, thank you for your support and for being in my life. I include in this category all my fellow colleagues in lab and collaborators at MIT – Euiheon Chung, Elaine Gee, Nikola Kojic, Paul Nerenberg, Christine Phillips, Ramon Salsas-Escat, Christian Schubert, and Veena Venkatachalam. As one of the first PhD students in Collin's group, my goal was to help establish a tone for the lab that was open, supportive, and collaborative. I am happy to see that these characteristics are now well integrated into the culture of our group. It has been a privilege to work alongside such a wonderful group of talented individuals. I am also grateful for all my friends and family outside MIT who have kept me grounded and connected to the world outside academia. I would not have gotten this far without your love and companionship. Finally, to Melissa, I say (as always), "te amo mucho."

References

[1] S. Hovmoller, T. Zhou, and T. Ohlson, "Conformations of amino acids in proteins," *Acta Crystallographica Section D-Biological Crystallography,* vol. 58, pp. 768-776, May 2002.

[2] S. Neal, A. M. Nip, H. Y. Zhang, and D. S. Wishart, "Rapid and accurate calculation of protein H-1, C-13 and N-15 chemical shifts," *Journal Of Biomolecular Nmr,* vol. 26, pp. 215-240, Jul 2003.

[3] J. L. Cummings, "Alzheimer's disease," *N Engl J Med,* vol. 351, pp. 56-67, Jul 1 2004.

[4] C. Mount and C. Downton, "Alzheimer disease: progress or profit?," *Nat Med,* vol. 12, pp. 780-4, Jul 2006.

[5] V. Kumar, Cotran, R.S., Robbins, S.L., *Robbins Pathology 7th ed.*, 7 ed. Philadelphia: W.B. Saunders Company, 2003.

[6] K. Eckermann, M. M. Mocanu, I. Khlistunova, J. Biernat, A. Nissen, A. Hofmann, K. Schonig, H. Bujard, A. Haemisch, E. Mandelkow, L. Zhou, G. Rune, and E. M. Mandelkow, "The beta-propensity of Tau determines aggregation and synaptic loss in inducible mouse models of tauopathy," *Journal Of Biological Chemistry,* vol. 282, pp. 31755-31765, Oct 26 2007.

[7] I. Khlistunova, J. Biernat, Y. P. Wang, M. Pickhardt, M. von Bergen, Z. Gazova, E. Mandelkow, and M. Mandelkow, "Inducible expression of tau repeat domain in cell models of tauopathy - Aggregation is toxic to cells but can be reversed by inhibitor drugs," *Journal Of Biological Chemistry,* vol. 281, pp. 1205-1214, Jan 13 2006.

[8] J. Lewis, D. W. Dickson, W. L. Lin, L. Chisholm, A. Corral, G. Jones, S. H. Yen, N. Sahara, L. Skipper, D. Yager, C. Eckman, J. Hardy, M. Hutton, and E. McGowan, "Enhanced neurofibrillary degeneration in transgenic mice expressing mutant tau and APP," *Science,* vol. 293, pp. 1487-1491, Aug 24 2001.

[9] D. J. Selkoe, "Alzheimer's disease is a synaptic failure," *Science,* vol. 298, pp. 789-91, Oct 25 2002.

[10] D. J. Selkoe and D. Schenk, "Alzheimer's disease: molecular understanding predicts amyloid-based therapeutics," *Annu Rev Pharmacol Toxicol,* vol. 43, pp. 545-84, 2003.

[11] A. Huang and Stultz CM., "Finding Order within Disorder - Elucidating the Structure of Proteins Associated with Neurodegenerative Disease," *Future Medicinal Chemistry (accepted),* 2009.

[12] P. V. Arriagada, J. H. Growdon, E. T. Hedleywhyte, and B. T. Hyman, "Neurofibrillary Tangles But Not Senile Plaques Parallel Duration And Severity Of Alzheimers-Disease," *Neurology,* vol. 42, pp. 631-639, Mar 1992.

[13] I. Khlistunova, M. Pickhardt, J. Biernat, Y. Wang, E. M. Mandelkow, and E. Mandelkow, "Inhibition of tau aggregation in cell models of tauopathy," *Curr Alzheimer Res,* vol. 4, pp. 544-6, Dec 2007.

[14] I. Grundke-Iqbal, K. Iqbal, Y. C. Tung, M. Quinlan, H. M. Wisniewski, and L. I. Binder, "Abnormal phosphorylation of the microtubule-associated protein tau (tau) in Alzheimer cytoskeletal pathology," *Proc Natl Acad Sci U S A,* vol. 83, pp. 4913-7, Jul 1986.

[15] J. M. Gulbis, Z. Kelman, J. Hurwitz, M. Odonnell, and J. Kuriyan, "Structure of the C-terminal region of p21(WAF1/CIP1) complexed with human PCNA," *Cell,* vol. 87, pp. 297-306, 1996.

[16] Y. H. Sung, J. Shin, J. H. Shin, and W. Lee, "Solution structure of p21(Waf1/Cip1/Sdi1) C-terminal domain bound to Cdk4," *Journal of Biomolecular Structure & Dynamics,* vol. 19, pp. 419-427, 2001.

[17] M.-K. Yoon, V. Venkatachalam, A. Huang, B.-S. Choi, C. Stultz, and J. Chou, "Residual structure within the disordered C-terminal segment of p21Waf1/Cip1/Sdi1 and its implications for molecular recognition," *Protein Science.,* vol. 18, pp. 337-347, 2009.

[18] E. Mylonas, A. Hascher, P. Bernado, M. Blackledge, E. Mandelkow, and D. I. Svergun, "Domain conformation of tau protein studied by solution small-angle X-ray scattering," *Biochemistry,* vol. 47, pp. 10345-10353, Sep 2008.

[19] M. D. Mukrasch, P. Markwick, J. Biernat, M. von Bergen, P. Bernado, C. Griesinger, E. Mandelkow, M. Zweckstetter, and M. Blackledge, "Highly populated turn conformations in natively unfolded Tau protein identified from residual dipolar couplings and molecular simulation," *Journal Of The American Chemical Society,* vol. 129, pp. 5235-5243, Apr 25 2007.

[20] J. L. F. Eric D. Scheeff, "Fundamentals of Protein Structure," in *Structural Bioinformatics*, H. W. Philip E. Bourne, Ed., 2005, pp. 15-39.

[21] A. K. Dunker, V. N. Uversky, C. J. Oldfield, A. Mohan, Y. Cheng, S. Zaidi, P. R. Romero, H. Xie, and Z. Obradovic, "Intrinsically disordered proteins," 2007, pp. 1A-1A.

[22] V. N. Uversky, "Natively unfolded proteins: A point where biology waits for physics," *Protein Science,* vol. 11, pp. 739-756, Apr 2002.

[23] V. N. Uversky, C. J. Oldfield, and A. K. Dunker, "Intrinsically disordered proteins in human diseases: Introducing the D-2 concept," *Annual Review of Biophysics,* vol. 37, pp. 215-246, 2008.

[24] M. Goedert and M. G. Spillantini, "A century of Alzheimer's disease," *Science,* vol. 314, pp. 777-81, Nov 3 2006.

[25] J. Marx, "Alzheimer's disease - A new take on tau," *Science,* vol. 316, pp. 1416-1417, Jun 2007.

[26] M. R. Cookson, "The biochemistry of Parkinson's disease," *Annual Review of Biochemistry,* vol. 74, pp. 29-52, 2005.

[27] R. Kruger, W. Kuhn, T. Muller, D. Woitalla, M. Graeber, S. Kosel, H. Przuntek, J. T. Epplen, L. Schols, and O. Riess, "Ala30Pro mutation in the gene encoding alpha-synuclein in Parkinson's disease," *Nature Genetics,* vol. 18, pp. 106-108, Feb 1998.

[28] M. Citron, T. Oltersdorf, C. Haass, L. McConlogue, A. Y. Hung, P. Seubert, C. Vigo-Pelfrey, I. Lieberburg, and D. J. Selkoe, "Mutation of the beta-amyloid precursor protein in familial Alzheimer's disease increases beta-protein production," *Nature,* vol. 360, pp. 672-4, Dec 17 1992.

[29] P. Rizzu, J. C. Van Swieten, M. Joosse, M. Hasegawa, M. Stevens, A. Tibben, M. F. Niermeijer, M. Hillebrand, R. Ravid, B. A. Oostra, M. Goedert, C. M. van Duijn, and P. Heutink, "High prevalence of mutations in the microtubule-associated protein tau in a population study of frontotemporal dementia in the Netherlands," *Am J Hum Genet,* vol. 64, pp. 414-21, Feb 1999.

[30] M. H. Polymeropoulos, C. Lavedan, E. Leroy, S. E. Ide, A. Dehejia, A. Dutra, B. Pike, H. Root, J. Rubenstein, R. Boyer, E. S. Stenroos, S. Chandrasekharappa, A. Athanassiadou, T. Papapetropoulos, W. G. Johnson, A. M. Lazzarini, R. C. Duvoisin, G. Dilorio, L. I. Golbe, and R. L. Nussbaum, "Mutation in the alpha-synuclein gene identified in families with Parkinson's disease," *Science,* vol. 276, pp. 2045-2047, Jun 27 1997.

[31] A. Goate, M.-C. Chartier-Harlin, M. Mullan, J. Brown, F. Crawford, L. Fidani, L. Giuffra, A. Haynes, N. Irving, L. James, R. Mant, P. Newton, K. Rooke, P. Roques, C. Talbot, M. Pericak-Vance, A. Roses, R. Williamson, M. Rossor, M. Owen, and J. Hardy, "Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease," *Nature,* vol. 349, pp. 704-706, 1991.

[32] K. P. S. J. Murphy, R. J. Carter, L. A. Lione, L. Mangiarini, A. Mahal, G. P. Bates, S. B. Dunnett, and A. J. Morton, "Abnormal Synaptic Plasticity and Impaired Spatial Cognition in Mice Transgenic for Exon 1 of the Human Huntington's Disease Mutation," *J. Neurosci.,* vol. 20, pp. 5115-5123, July 1, 2000 2000.

[33] B. L. Schneider, C. R. Seehus, E. E. Capowski, P. Aebischer, S. C. Zhang, and C. N. Svendsen, "Over-expression of alpha-synuclein in human neural progenitors leads to specific changes in fate and differentiation," *Hum Mol Genet,* vol. 16, pp. 651-66, Mar 15 2007.

[34] M. Pickhardt, Z. Gazova, M. von Bergen, I. Khlistunova, Y. Wang, A. Hascher, E. M. Mandelkow, J. Biernat, and E. Mandelkow, "Anthraquinones inhibit tau aggregation and dissolve Alzheimer's paired helical filaments in vitro and in cells," *J Biol Chem,* vol. 280, pp. 3628-35, Feb 4 2005.

[35] J. Ghanta, C. L. Shen, L. L. Kiessling, and R. M. Murphy, "A strategy for designing inhibitors of beta-amyloid toxicity," *J Biol Chem,* vol. 271, pp. 29525-8, Nov 22 1996.

[36] V. N. Uversky and A. L. Fink, "Conformational constraints for amyloid fibrillation: the importance of being unfolded," *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics,* vol. 1698, pp. 131-153, 2004.

[37] V. N. Uversky, "Amyloidogenesis of natively unfolded proteins," *Current Alzheimer Research,* vol. 5, pp. 260-287, Jun 2008.

[38] A. Vitalis, X. L. Wang, and R. V. Pappu, "Quantitative characterization of intrinsic disorder in polyglutamine: Insights from analysis based on polymer theories," *Biophysical Journal,* vol. 93, pp. 1923-1937, Sep 2007.

[39] A. Vitalis, X. L. Wang, and R. V. Pappu, "Atomistic Simulations of the Effects of Polyglutamine Chain Length and Solvent Quality on Conformational Equilibria and Spontaneous Homodimerization," *Journal Of Molecular Biology,* vol. 384, pp. 279-297, Dec 2008.

[40] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White, Jr., "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain," *Proc Natl Acad Sci U S A,* vol. 47, pp. 1309-14, Sep 15 1961.

[41] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips, "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis," *Nature,* vol. 181, pp. 662-6, Mar 8 1958.

[42] L. Bragg and M. F. Perutz, "THE STRUCTURE OF HAEMOGLOBIN .6. FOURIER PROJECTIONS ON THE 010-PLANE," *Proceedings of the Royal Society of London Series a-Mathematical and Physical Sciences,* vol. 225, pp. 315-329, 1954.

[43] P. J. Flory, *Statistical Mechanics of Chain Molecules*. New York: Interscience Publishers, 1969.

[44] J. D. Bryngelson and P. G. Wolynes, "Spin-Glasses and the Statistical-Mechanics of Protein Folding," *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 84, pp. 7524-7528, Nov 1987.

[45] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, "Funnels, Pathways, and the Energy Landscape of Protein-Folding - a Synthesis," *Proteins-Structure Function And Genetics,* vol. 21, pp. 167-195, Mar 1995.

[46] Levinthal, "How to Fold Graciously," *Mössbaun Spectroscopy in Biological Systems Proceedings,* vol. 67, pp. 22-24, 1969.

[47] T. Lazaridis and M. Karplus, "Thermodynamics of protein folding: a microscopic view," *Biophys Chem,* vol. 100, pp. 367-95, 2003.

[48] A. Huang and C. M. Stultz, "Conformational sampling with implicit solvent models: Application to the PHF6 peptide in tau protein," *Biophysical Journal,* vol. 92, pp. 34-45, Jan 2007.

[49] A. Huang and C. M. Stultz, "The Effect of a ΔK280 Mutation on the Unfolded State of a Microtubule Binding Repeat in Tau," *PLoS Computational Biology,* vol. 4(8): e1000155, pp. 1-12, 2008.

[50] A. K. Dunker, I. Silman, V. N. Uversky, and J. L. Sussman, "Function and structure of inherently disordered proteins," *Curr Opin Struct Biol,* vol. 18, pp. 756-64, Dec 2008.

[51] V. Receveur-Brechot, J. M. Bourhis, V. N. Uversky, B. Canard, and S. Longhi, "Assessing protein disorder and induced folding," *Proteins-Structure Function and Bioinformatics,* vol. 62, pp. 24-45, Jan 1 2006.

[52] T. Mittag and J. D. Forman-Kay, "Atomic-level characterization of disordered protein ensembles," *Current Opinion In Structural Biology,* vol. 17, pp. 3-14, Feb 2007.

[53] D. Eliezer, "Biophysical characterization of intrinsically disordered proteins," *Curr Opin Struct Biol,* vol. 19, pp. 23-30, Feb 2009.

[54] F. Zhu, J. Kapitan, G. E. Tranter, P. D. A. Pudney, N. W. Isaacs, L. Hecht, and L. D. Barron, "Residual structure in disordered peptides and unfolded proteins from multivariate analysis and ab initio simulation of Raman optical activity data," *Proteins-Structure Function And Bioinformatics,* vol. 70, pp. 823-833, Feb 2008.

[55] O. Schweers, E. Schonbrunnhanebeck, A. Marx, and E. Mandelkow, "Structural Studies Of Tau-Protein And Alzheimer Paired Helical Filaments Show No Evidence For Beta-Structure," *Journal Of Biological Chemistry,* vol. 269, pp. 24290-24297, Sep 30 1994.

[56] K. Wuthrich, "NMR - this other method for protein and nucleic acid structure determination," *Acta Crystallogr D Biol Crystallogr,* vol. 51, pp. 249-70, May 1 1995.

[57] M. M. Dedmon, K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo, and C. M. Dobson, "Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations," *Journal Of The American Chemical Society,* vol. 127, pp. 476-477, Jan 19 2005.

[58] P. Bernado, C. W. Bertoncini, C. Griesinger, M. Zweckstetter, and M. Blackledge, "Defining Long-Range Order and Local Disorder in Native alpha-Synuclein Using Residual Dipolar Couplings," *Journal Of The American Chemical Society,* vol. 127, pp. 17968-17969, 2005.

[59] S. Meier, M. Blackledge, and S. Grzesiek, "Conformational distributions of unfolded polypeptides from novel NMR techniques," *Journal Of Chemical Physics,* vol. 128, Feb 2008.

[60] M. Louhivuori, K. Fredriksson, K. Paakkonen, P. Permi, and A. Annila, "Alignment of chain-like molecules," *Journal Of Biomolecular Nmr,* vol. 29, pp. 517-524, Aug 2004.

[61] C. T. Jiang and J. Y. Chang, "Isomers of human alpha-synuclein stabilized by disulfide bonds exhibit distinct structural and aggregative properties," *Biochemistry,* vol. 46, pp. 602-609, Jan 16 2007.

[62] J. Y. Chang and L. Li, "The structure of denatured alpha-lactalbumin elucidated by the technique of disulfide scrambling - Fractionation of conformational isomers of alpha-lactalbumin," *Journal Of Biological Chemistry,* vol. 276, pp. 9705-9712, Mar 30 2001.

[63] G. J. Arlaud, P. N. Barlow, C. Gaboriaud, P. Gros, and S. V. L. Narayana, "Deciphering complement mechanisms: The contributions of structural biology," 2007, pp. 3809-3822.

[64] R. A. S. Ariens, T. S. Lai, J. W. Weisel, C. S. Greenberg, and P. J. Grant, "Role of factor XIII in fibrin clot formation and effects of genetic polymorphisms," *Blood,* vol. 100, pp. 743-754, Aug 2002.

[65] V. N. Uversky, "A protein-chameleon: Conformational plasticity of alpha-synuclein, a disordered protein involved in neurodegenerative disorders," *Journal of Biomolecular Structure & Dynamics,* vol. 21, pp. 211-234, Oct 2003.

[66] M. Sandal, F. Valle, I. Tessari, S. Mammi, E. Bergantino, F. Musiani, M. Brucale, L. Bubacco, and B. Samori, "Conformational equilibria in monomeric alpha-synuclein at the single-molecule level," *PLoS Biol,* vol. 6, p. e6, Jan 2008.

[67] A. K. Jha, A. Colubri, K. F. Freed, and T. R. Sosnick, "Statistical coil model of the unfolded state: Resolving the reconciliation problem," *Proceedings of the National Academy of Sciences,* vol. 102, p. 13099, 2005.

[68] P. Bernado, L. Blanchard, P. Timmins, D. Marion, R. W. H. Ruigrok, and M. Blackledge, "A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering," *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 102, pp. 17002-17007, Nov 22 2005.

[69] J. Li, V. N. Uversky, and A. L. Fink, "Effect of familial Parkinson's disease point mutations A30P and A53T on the structural properties, aggregation, and fibrillation of human alpha-synuclein," *Biochemistry,* vol. 40, pp. 11604-11613, Sep 25 2001.

[70] R. Bussell and D. Eliezer, "Residual structure and dynamics in Parkinson's disease-associated mutants of alpha-synuclein," *Journal Of Biological Chemistry,* vol. 276, pp. 45996-46003, Dec 7 2001.

[71] K. A. Conway, S. J. Lee, J. C. Rochet, T. T. Ding, R. E. Williamson, and P. T. Lansbury, "Acceleration of oligomerization, not fibrillization, is a shared property of both alpha-synuclein mutations linked to early-onset Parkinson's disease: Implications for pathogenesis and therapy," *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 97, pp. 571-576, Jan 18 2000.

[72] C. W. Bertoncini, C. O. Fernandez, C. Griesinger, T. M. Jovin, and M. Zweckstetter, "Familial mutants of alpha-synuclein with increased neurotoxicity have a destabilized conformation," *Journal Of Biological Chemistry,* vol. 280, pp. 30649-30652, Sep 2 2005.

[73] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chemical Physics Letters,* vol. 314, pp. 141-151, 1999.

[74] A. Baumketner, S. L. Bernstein, T. Wyttenbach, N. D. Lazo, D. B. Teplow, M. T. Bowers, and J. E. Shea, "Structure of the 21-30 fragment of amyloid beta-protein," *Protein Science,* vol. 15, pp. 1239-1247, Jun 2006.

[75] X. Wang, A. Vitalis, M. A. Wyczalkowski, and R. V. Pappu, "Characterizing the conformational ensemble of monomeric polyglutamine," *Proteins,* vol. 63, pp. 297-311, May 1 2006.

[76] M. Vendruscolo, "Determination of conformationally heterogeneous states of proteins," *Current Opinion In Structural Biology,* vol. 17, pp. 15-20, Feb 2007.

[77] W. Y. Choy and J. D. Forman-Kay, "Calculation of ensembles of structures representing the unfolded state of an SH3 domain," *Journal Of Molecular Biology,* vol. 308, pp. 1011-1032, May 18 2001.

[78] J. A. Marsh, C. Neale, F. E. Jack, W. Y. Choy, A. Y. Lee, K. A. Crowhurst, and J. D. Forman-Kay, "Improved structural characterizations of the drkN SH3 domain unfolded state suggest a compact ensemble with native-like and non-native structure," *Journal Of Molecular Biology,* vol. 367, pp. 1494-1510, Apr 13 2007.

[79] Y. Chen, S. L. Campbell, and N. V. Dokholyan, "Deciphering protein dynamics from NMR data using explicit structure sampling and selection," *Biophysical Journal,* vol. 93, pp. 2300-2306, Oct 2007.

[80] P. Bernado, E. Mylonas, M. V. Petoukhov, M. Blackledge, and D. I. Svergun, "Structural characterization of flexible proteins using small-angle X-ray scattering," *Journal Of The American Chemical Society,* vol. 129, pp. 5656-5664, May 2007.

[81] K. Osapay and D. A. Case, "A new analysis of proton chemical shifts in proteins," *Journal Of The American Chemical Society,* vol. 113, pp. 9436-9444, 1991.

[82] M. Zweckstetter and A. Bax, "Prediction of sterically induced alignment in a dilute liquid crystalline phase: Aid to protein structure determination by NMR," *Journal of the American Chemical Society,* vol. 122, pp. 3791-3792, Apr 2000.

[83] C. L. Masters, G. Simms, N. A. Weinman, G. Multhaup, B. L. McDonald, and K. Beyreuther, "Amyloid plaque core protein in Alzheimer disease and Down syndrome," *Proc Natl Acad Sci U S A,* vol. 82, pp. 4245-9, Jun 1985.

[84] K. S. Kosik, L. D. Orecchio, L. Binder, J. Q. Trojanowski, V. M. Lee, and G. Lee, "Epitopes that span the tau molecule are shared with paired helical filaments," *Neuron,* vol. 1, pp. 817-25, Nov 1988.

[85] J. Hardy and D. J. Selkoe, "The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics," *Science,* vol. 297, pp. 353-6, Jul 19 2002.

[86] M. Coles, W. Bicknell, A. A. Watson, D. P. Fairlie, and D. J. Craik, "Solution structure of amyloid beta-peptide(1-40) in a water-micelle environment. Is the membrane-spanning domain where we think it is?," *Biochemistry,* vol. 37, pp. 11064-77, Aug 4 1998.

[87] S. Zhang, K. Iwata, M. J. Lachenmann, J. W. Peng, S. Li, E. R. Stimson, Y. Lu, A. M. Felix, J. E. Maggio, and J. P. Lee, "The Alzheimer's peptide A beta adopts a collapsed coil structure in water," *Journal of Structural Biology,* vol. 130, pp. 130-141, Jun 2000.

[88] L. M. Hou, H. Y. Shao, Y. B. Zhang, H. Li, N. K. Menon, E. B. Neuhaus, J. M. Brewer, I. J. L. Byeon, D. G. Ray, M. P. Vitek, T. Iwashita, R. A. Makula, A. B. Przybyla, and M. G. Zagorski, "Solution NMR studies of the A beta(1-40) and A beta(1-42) peptides establish that the met35 oxidation state affects the mechanism of amyloid formation," *Journal Of The American Chemical Society,* vol. 126, pp. 1992-2005, Feb 2004.

[89] A. Baumketner and J. E. Shea, "The structure of the Alzheimer amyloid beta 10-35 peptide probed through replica-exchange molecular dynamics simulations in explicit solvent," *J Mol Biol,* vol. 366, pp. 275-85, Feb 9 2007.

[90] J. Khandogin and C. L. Brooks, "Linking folding with aggregation in Alzheimer's beta-amyloid peptides," *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 104, pp. 16880-16885, Oct 2007.

[91] B. Urbanc, L. Cruz, F. Ding, D. Sammond, S. Khare, S. V. Buldyrev, H. E. Stanley, and N. V. Dokholyan, "Molecular dynamics simulation of amyloid beta dimer formation," *Biophys J,* vol. 87, pp. 2310-21, Oct 2004.

[92] P. H. Nguyen, M. S. Li, G. Stock, J. E. Straub, and D. Thirumalai, "Monomer adds to preformed structured oligomers of Abeta-peptides by a two-stage dock-lock mechanism," *Proc Natl Acad Sci U S A,* vol. 104, pp. 111-6, Jan 2 2007.

[93] N. L. Fawzi, K. L. Kohlstedt, Y. Okabe, and T. Head-Gordon, "Protofibril assemblies of the arctic, Dutch, and Flemish mutants of the Alzheimer's Abeta1-40 peptide," *Biophys J,* vol. 94, pp. 2007-16, Mar 15 2008.

[94] W. Hwang, S. Zhang, R. D. Kamm, and M. Karplus, "Kinetic control of dimer structure formation in amyloid fibrillogenesis," *Proc Natl Acad Sci U S A,* vol. 101, pp. 12916-21, Aug 31 2004.

[95] S. Gnanakaran, R. Nussinov, and A. E. Garcia, "Atomic-level description of amyloid beta-dimer formation," *J Am Chem Soc,* vol. 128, pp. 2158-9, Feb 22 2006.

[96] N. V. Buchete, R. Tycko, and G. Hummer, "Molecular dynamics simulations of Alzheimer's beta-amyloid protofilaments," *J Mol Biol,* vol. 353, pp. 804-21, Nov 4 2005.

[97] C. W. Olanow and W. G. Tatton, "Etiology and pathogenesis of Parkinson's disease," *Annual Review Of Neuroscience,* vol. 22, pp. 123-144, 1999.

[98] W. R. Gibb and A. J. Lees, "The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease," *J Neurol Neurosurg Psychiatry,* vol. 51, pp. 745-52, Jun 1988.

[99] P. H. Weinreb, W. G. Zhen, A. W. Poon, K. A. Conway, and P. T. Lansbury, "NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded," *Biochemistry,* vol. 35, pp. 13709-13715, Oct 29 1996.

[100] D. F. Clayton and J. M. George, "Synucleins in synaptic plasticity and neurodegenerative disorders," *J Neurosci Res,* vol. 58, pp. 120-9, Oct 1 1999.

[101] M. L. Hegde and K. S. J. Rao, "Challenges and complexities of alpha-synuclein toxicity: new postulates in unfolding the mystery associated with Parkinson's disease," *Archives of Biochemistry and Biophysics,* vol. 418, pp. 169-178, Oct 15 2003.

[102] H. Y. Han, P. H. Weinreb, and P. T. Lansbury, "The Core Alzheimers Peptide Nac Forms Amyloid Fibrils Which Seed and Are Seeded by Beta-Amyloid - Is Nac a Common Trigger or Target in Neurodegenerative Disease," *Chemistry & Biology,* vol. 2, pp. 163-169, Mar 1995.

[103] K. Ueda, H. Fukushima, E. Masliah, Y. Xia, A. Iwai, M. Yoshimoto, D. A. C. Otero, J. Kondo, Y. Ihara, and T. Saitoh, "MOLECULAR-CLONING OF CDNA-ENCODING AN UNRECOGNIZED COMPONENT OF AMYLOID IN ALZHEIMER-DISEASE," *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 90, pp. 11282-11286, Dec 1993.

[104] A. Iwai, E. Masliah, M. Yoshimoto, N. Ge, L. Flanagan, H. A. Rohan de Silva, A. Kittel, and T. Saitoh, "The precursor protein of non-A[beta] component of Alzheimer's disease amyloid is a presynaptic protein of the central nervous system," *Neuron,* vol. 14, pp. 467-475, 1995.

[105] F. Chiti and C. M. Dobson, "Protein misfolding, functional amyloid, and human disease," *Annual Review of Biochemistry,* vol. 75, pp. 333-366, 2006.

[106] C. A. Ross and M. A. Poirier, "Protein aggregation and neurodegenerative disease," *Nature Medicine,* pp. S10-S17, Jul 2004.

[107] V. N. Uversky, J. Li, and A. L. Fink, "Evidence for a partially folded intermediate in alpha-synuclein fibril formation," *Journal Of Biological Chemistry,* vol. 276, pp. 10737-10744, Apr 6 2001.

[108] A. Der-Sarkissian, C. C. Jao, J. Chen, and R. Langen, "Structural organization of alpha-synuclein fibrils studied by site-directed spin labeling," *Journal Of Biological Chemistry,* vol. 278, pp. 37530-37535, Sep 26 2003.

[109] M. Vilar, H. T. Chou, T. Luhrs, S. K. Maji, D. Riek-Loher, R. Verel, G. Manning, H. Stahlberg, and R. Riek, "The fold of alpha-synuclein fibrils," *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 105, pp. 8637-8642, Jun 24 2008.

[110] N. P. Ulrih, C. H. Barry, and A. L. Fink, "Impact of Tyr to Ala mutations on alpha-synuclein fibrillation and structural properties," *Biochimica Et Biophysica Acta-Molecular Basis Of Disease,* vol. 1782, pp. 581-585, Oct 2008.

[111] C. W. Bertoncini, Y. S. Jung, C. O. Fernandez, W. Hoyer, C. Griesinger, T. M. Jovin, and M. Zweckstetter, "Release of long-range tertiary interactions potentiates aggregation of natively unstructured alpha-synuclein," *Proc Natl Acad Sci U S A,* vol. 102, pp. 1430-5, Feb 1 2005.

[112] A. Andreadis, "Tau gene alternative splicing: expression patterns, regulation and modulation of function in normal brain and neurodegenerative diseases," *Biochimica Et Biophysica Acta-Molecular Basis Of Disease,* vol. 1739, pp. 91-103, Jan 3 2005.

[113] M. Goedert, "Tau gene mutations and their effects," *Movement Disorders,* vol. 20, pp. S45-S52, Aug 2005.

[114] K. Iqbal, A. D. C. Alonso, S. Chen, M. O. Chohan, E. El-Akkad, C. X. Gong, S. Khatoon, B. Li, F. Liu, A. Rahman, H. Tanimukai, and I. Grundke-Iqbal, "Tau pathology in Alzheimer disease and other tauopathies," *Biochimica Et Biophysica Acta-Molecular Basis Of Disease,* vol. 1739, pp. 198-210, Jan 3 2005.

[115] K. S. Kosik and H. Shimura, "Phosphorylated tau and the neurodegenerative foldopathies," *Biochimica Et Biophysica Acta-Molecular Basis Of Disease,* vol. 1739, pp. 298-310, Jan 3 2005.

[116] B. L. Goode, P. E. Denis, D. Panda, M. J. Radeke, H. P. Miller, L. Wilson, and S. C. Feinstein, "Functional interactions between the proline-rich and repeat regions of tau enhance microtubule binding and assembly," *Molecular Biology of the Cell,* vol. 8, pp. 353-365, Feb 1997.

[117] M. Necula and J. Kuret, "Pseudophosphorylation and glycation of tau protein enhance but do not trigger fibrillization in vitro," *Journal Of Biological Chemistry,* vol. 279, pp. 49694-49703, Nov 26 2004.

[118] M. Necula and J. Kuret, "Site-specific pseudophosphorylation modulates the rate of tau filament dissociation," *Febs Letters,* vol. 579, pp. 1453-1457, Feb 28 2005.

[119] N. R. Graff-Radford and B. K. Woodruff, "Frontotemporal dementia," *Seminars in Neurology,* vol. 27, pp. 48-57, Feb 2007.

[120] A. D. Alonso, A. Mederlyova, M. Novak, I. Grundke-Iqbal, and K. Iqbal, "Promotion of hyperphosphorylation by frontotemporal dementia tau mutations," *Journal Of Biological Chemistry,* vol. 279, pp. 34873-34881, Aug 13 2004.

[121] S. Barghorn, Q. Zheng-Fischhofer, M. Ackmann, J. Biernat, M. von Bergen, E. M. Mandelkow, and E. Mandelkow, "Structure, microtubule interactions, and paired helical filament aggregation by

tau mutants of frontotemporal dementias," *Biochemistry,* vol. 39, pp. 11714-11721, Sep 26 2000.

[122] D. Eliezer, P. Barre, M. Kobaslija, D. Chan, X. H. Li, and L. Heend, "Residual structure in the repeat domain of tau: Echoes of microtubule binding and paired helical filament formation," *Biochemistry,* vol. 44, pp. 1026-1036, Jan 25 2005.

[123] M. D. Mukrasch, J. Biernat, M. von Bergen, C. Griesinger, E. Mandelkow, and M. Zweckstetter, "Sites of tau important for aggregation populate beta-structure and bind to microtubules and polyanions," *Journal Of Biological Chemistry,* vol. 280, pp. 24978-24986, Jul 1 2005.

[124] D. Fischer, M. D. Mukrasch, M. von Bergen, A. Klos-Witkowska, J. Biernat, C. Griesinger, E. Mandelkow, and M. Zweckstetter, "Structural and microtubule binding properties of tau mutants of frontotemporal dementias," *Biochemistry,* vol. 46, pp. 2574-2582, Mar 13 2007.

[125] I. Melnikova, "Therapies for Alzheimer's disease," *Nat Rev Drug Discov,* vol. 6, pp. 341-2, May 2007.

[126] B. Bulic, M. Pickhardt, I. Khlistunova, J. Biernat, E. M. Mandelkow, E. Mandelkow, and H. Waldmann, "Rhodanine-based tau aggregation inhibitors in cell models of tauopathy," *Angew Chem Int Ed Engl,* vol. 46, pp. 9215-9, 2007.

[127] C. M. Wischik, P. C. Edwards, R. Y. Lai, M. Roth, and C. R. Harrington, "Selective inhibition of Alzheimer disease-like tau aggregation by phenothiazines," *Proc Natl Acad Sci U S A,* vol. 93, pp. 11213-8, Oct 1 1996.

[128] J. Greer, J. W. Erickson, J. J. Baldwin, and M. D. Varney, "Application of the three-dimensional structures of protein target molecules in structure-based drug design," *J Med Chem,* vol. 37, pp. 1035-54, Apr 15 1994.

[129] Y. Cheng, T. LeGall, C. J. Oldfield, J. P. Mueller, Y. Y. Van, P. Romero, M. S. Cortese, V. N. Uversky, and A. K. Dunker, "Rational drug design via intrinsically disordered protein," *Trends Biotechnol,* vol. 24, pp. 435-42, Oct 2006.

[130] C. M. Stultz and M. Karplus, "Fragment-based Approaches in Drug Discovery." vol. 34, 2007, pp. 125-148.

[131] A. C. Anderson, "The process of structure-based drug design," *Chem Biol,* vol. 10, pp. 787-97, Sep 2003.

[132] M. Feig and C. L. Brooks, 3rd, "Recent advances in the development and application of implicit solvent models in biomolecule simulations," *Curr Opin Struct Biol,* vol. 14, pp. 217-24, Apr 2004.

[133] B. Roux and T. Simonson, "Implicit solvent models," *Biophys Chem,* vol. 78, pp. 1-20, Apr 5 1999.

[134] C. L. Brooks, 3rd and M. Karplus, "Solvent effects on protein motion and protein effects on solvent motion. Dynamics of the active site region of lysozyme," *J Mol Biol,* vol. 208, pp. 159-81, Jul 5 1989.

[135] A. Jaramillo and S. J. Wodak, "Computational protein design is a challenge for implicit solvation models," *Biophys J,* vol. 88, pp. 156-71, Jan 2005.

[136] R. Zhou and B. J. Berne, "Can a continuum solvent model reproduce the free energy landscape of a beta -hairpin folding in water?," *Proc Natl Acad Sci U S A,* vol. 99, pp. 12777-82, Oct 1 2002.

[137] C. M. Stultz, "An assessment of potential of mean force calculations with implicit solvent models," *Journal Of Physical Chemistry B,* vol. 108, pp. 16525-16532, Oct 21 2004.

[138] B. N. Dominy and C. L. Brooks, "Development of a generalized born model parametrization for proteins and nucleic acids," *Journal Of Physical Chemistry B,* vol. 103, pp. 3765-3773, May 1999.

[139] W. Im, M. S. Lee, and C. L. Brooks, 3rd, "Generalized born model with a simple smoothing function," *J Comput Chem,* vol. 24, pp. 1691-702, Nov 15 2003.

[140] T. Lazaridis and M. Karplus, "Effective energy function for proteins in solution," *Proteins-Structure Function And Genetics,* vol. 35, pp. 133-152, May 1 1999.

[141] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of Simple Potential Functions for Simulating Liquid Water," *Journal Of Chemical Physics,* vol. 79, pp. 926-935, 1983.

[142] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, "Semilanalytical Treatment of Solvation for Molecular Mechanics and Dynamics," *Journal Of The American Chemical Society,* vol. 112, pp. 6127-6129, Aug 1990.

[143] T. H. Rod, J. L. Radkiewicz, and C. L. Brooks, 3rd, "Correlated motion and the effect of distal mutations in dihydrofolate reductase," *Proc Natl Acad Sci U S A,* vol. 100, pp. 6980-5, Jun 10 2003.

[144] J. Chen, H. S. Won, W. Im, H. J. Dyson, and C. L. Brooks, 3rd, "Generation of native-like protein structures from limited NMR data, modern force fields and advanced conformational sampling," *J Biomol NMR,* vol. 31, pp. 59-64, Jan 2005.

[145] Y. Liu, M. Scolari, W. Im, and H. J. Woo, "Protein-protein interactions in actin-myosin binding and structural effects of R405Q mutation: a molecular dynamics study," *Proteins,* vol. 64, pp. 156-66, Jul 1 2006.

[146] T. Lazaridis and M. Karplus, ""New view" of protein folding reconciled with the old through multiple unfolding simulations," *Science,* vol. 278, pp. 1928-31, Dec 12 1997.

[147] T. Lazaridis and M. Karplus, "Discrimination of the native from misfolded protein models with an energy function including implicit solvation," *Journal Of Molecular Biology,* vol. 288, pp. 477-487, 1999.

[148] Y. Inuzuka and T. Lazaridis, "On the unfolding of alpha-lytic protease and the role of the pro region," *Proteins: Structure, Function, and Genetics,* vol. 41, pp. 21-32, 2000.

[149] R. E. Bruccoleri and M. Karplus, "Conformational Sampling Using High-Temperature Molecular-Dynamics," *Biopolymers,* vol. 29, pp. 1847-1862, Dec 1990.

[150] C. M. Stultz and M. Karplus, "MCSS functionality maps for a flexible protein," *Proteins-Structure Function And Genetics,* vol. 37, pp. 512-529, Dec 1 1999.

[151] A. Ashish and R. Kishore, "Folded conformation of an immunostimulating tetrapeptide Rigin: high temperature molecular dynamics simulation study," *Bioorganic & Medicinal Chemistry,* vol. 10, pp. 4083-4090, 2002.

[152] S. D. O'Connor, P. E. Smith, F. Al-Obeidi, and B. M. Pettitt, "Quenched molecular dynamics simulations of tuftsin and proposed cyclic analogs," *Journal of Medicinal Chemistry,* vol. 35, pp. 2870-2881, 1992.

[153] D. C. Sullivan and C. Lim, "Toward absolute density of states calculations for proteins," *Journal Of Physical Chemistry B,* vol. 110, pp. 12125-12128, Jun 22 2006.

[154] M. von Bergen, P. Friedhoff, J. Biernat, J. Heberle, E. M. Mandelkow, and E. Mandelkow, "Assembly of tau protein into Alzheimer paired helical filaments depends on a local sequence motif ((306)VQIVYK(311)) forming beta structure," *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 97, pp. 5129-5134, May 9 2000.

[155] M. von Bergen, S. Barghorn, L. Li, A. Marx, J. Biernat, E. M. Mandelkow, and E. Mandelkow, "Mutations of tau protein in frontotemporal dementia promote aggregation of paired helical filaments by enhancing local beta-structure," *Journal Of Biological Chemistry,* vol. 276, pp. 48165-48174, Dec 21 2001.

[156] T. C. Gamblin, "Potential structure/function relationships of predicted secondary structural elements of tau," *Biochim Biophys Acta,* vol. 1739, pp. 140-9, Jan 3 2005.

[157] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "Charmm - A Program For Macromolecular Energy, Minimization, And Dynamics Calculations," *Journal Of Computational Chemistry,* vol. 4, pp. 187-217, 1983.

[158] C. L. Brooks, Karplus, M., Petitt, B.M., *Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*. New York: John Wiley and Sons, 1988.

[159] H. J. C. Berendsen, J. P. M. Postma, W. F. Vangunsteren, A. Dinola, and J. R. Haak, "Molecular-Dynamics With Coupling To An External Bath," *Journal Of Chemical Physics,* vol. 81, pp. 3684-3690, 1984.

[160] W. F. Vangunsteren and H. J. C. Berendsen, "Algorithms For Macromolecular Dynamics And Constraint Dynamics," *Molecular Physics,* vol. 34, pp. 1311-1327, 1977.

[161] M. Y. Shen and K. F. Freed, "Long time dynamics of Met-enkephalin: comparison of explicit and implicit solvent models," *Biophys J,* vol. 82, pp. 1791-808, Apr 2002.

[162] B. D. Bursulaya, M. Totrov, R. Abagyan, and C. L. Brooks, 3rd, "Comparative study of several algorithms for flexible ligand docking," *J Comput Aided Mol Des,* vol. 17, pp. 755-63, Nov 2003.

[163] W. Humphrey, A. Dalke, and K. Schulten, "VMD: Visual molecular dynamics," *Journal Of Molecular Graphics,* vol. 14, pp. 33-&, Feb 1996.

[164] C. M. Stultz, "Localized unfolding of collagen explains collagenase cleavage near imino-poor sites," *J Mol Biol,* vol. 319, pp. 997-1003, Jun 21 2002.

[165] B. Tidor and M. Karplus, "The contribution of vibrational entropy to molecular association. The dimerization of insulin," *J Mol Biol,* vol. 238, pp. 405-14, May 6 1994.

[166] D. Mcquarrie, *Statistical Mechanics*. Sausalito: University Science Books, 2000.

[167] B. N. Dominy and C. L. Brooks, "Identifying native-like protein structures using physics-based potentials," *J Comput Chem,* vol. 23, pp. 147-60, Jan 15 2002.

[168] M. C. Lee and Y. Duan, "Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model," *Proteins,* vol. 55, pp. 620-34, May 15 2004.

[169] A. K. Felts, E. Gallicchio, A. Wallqvist, and R. M. Levy, "Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the Surface Generalized Born solvent model," *Proteins,* vol. 48, pp. 404-22, Aug 1 2002.

[170] M. R. Lee, Y. Duan, and P. A. Kollman, "Use of MM-PB/SA in estimating the free energies of proteins: application to native, intermediates, and unfolded villin headpiece," *Proteins,* vol. 39, pp. 309-16, Jun 1 2000.

[171] Y. N. Vorobjev, J. C. Almagro, and J. Hermans, "Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum model," *Proteins,* vol. 32, pp. 399-413, Sep 1 1998.

[172] A. Masunov and T. Lazaridis, "Potentials of mean force between ionizable amino acid side chains in water," *J Am Chem Soc,* vol. 125, pp. 1722-30, Feb 19 2003.

[173] C. M. Dobson, "Principles of protein folding, misfolding and aggregation," *Seminars In Cell & Developmental Biology,* vol. 15, pp. 3-16, Feb 2004.

[174] M. Margittai and R. Langen, "Template-assisted filament growth by parallel stacking of tau," *Proc Natl Acad Sci U S A,* vol. 101, pp. 10278-83, Jul 13 2004.

[175] J. Berriman, L. C. Serpell, K. A. Oberg, A. L. Fink, M. Goedert, and R. A. Crowther, "Tau filaments from human brain and from in vitro assembly of recombinant protein show cross-beta structure," *Proc Natl Acad Sci U S A,* vol. 100, pp. 9034-8, Jul 22 2003.

[176] M. Tolnay and A. Probst, "The neuropathological spectrum of neurodegenerative tauopathies," *Iubmb Life,* vol. 55, pp. 299-305, Jun 2003.

[177] D. J. Selkoe, "Cell biology of protein misfolding: The examples of Alzheimer's and Parkinson's diseases," *Nature Cell Biology,* vol. 6, pp. 1054-1061, Nov 2004.

[178] G. Drewes, A. Ebneth, and E. M. Mandelkow, "MAPs, MARKs and microtubule dynamics," *Trends In Biochemical Sciences,* vol. 23, pp. 307-311, Aug 1998.

[179] M. D. Weingarten, A. H. Lockwood, S. Y. Hwo, and M. W. Kirschner, "Protein Factor Essential For Microtubule Assembly," *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 72, pp. 1858-1862, 1975.

[180] N. C. Fitzkee and G. D. Rose, "Reassessing random-coil statistics in unfolded proteins," *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 101, pp. 12497-12502, Aug 24 2004.

[181] S. W. Carlson, M. Branden, K. Voss, Q. Sun, C. A. Rankin, and T. C. Gamblin, "A complex mechanism for inducer mediated tau polymerization," *Biochemistry,* vol. 46, pp. 8838-8849, Jul 31 2007.

[182] M. Goedert and R. Jakes, "Mutations causing neurodegenerative tauopathies," *Biochimica Et Biophysica Acta-Molecular Basis Of Disease,* vol. 1739, pp. 240-250, Jan 3 2005.

[183] P. Rizzu, J. C. Van Swieten, M. Joosse, M. Hasegawa, M. Stevens, A. Tibben, M. F. Niermeijer, M. Hillebrand, R. Ravid, B. A. Oostra, M. Goedert, C. M. van Duijn, and P. Heutink, "High prevalence of mutations in the microtubule-associated protein tau in a population study of frontotemporal dementia in the Netherlands," *American Journal Of Human Genetics,* vol. 64, pp. 414-421, Feb 1999.

[184] J. C. van Swieten, M. Stevens, S. M. Rosso, P. Rizzu, M. Joosse, I. de Koning, W. Kamphorst, R. Ravid, M. G. Spillantini, M. F. Niermeijer, and P. Heutink, "Phenotypic variation in hereditary frontotemporal dementia with tau mutations," *Annals Of Neurology,* vol. 46, pp. 617-626, Oct 1999.

[185] W. Fieber, S. Kristjansdottir, and F. M. Poulsen, "Short-range, long-range and transition state interactions in the denatured state of ACBP from residual dipolar couplings," *Journal Of Molecular Biology,* vol. 339, pp. 1191-1199, Jun 18 2004.

[186] S. Meier, S. Guthe, T. Kiefhaber, and S. Grzesiek, "Foldon, the natural trimerization domain of T4 fibritin, dissociates into a monomeric A-state form containing a stable beta-hairpin: Atomic details of trimer dissociation and local beta-hairpin stability from residual dipolar couplings," *Journal Of Molecular Biology,* vol. 344, pp. 1051-1069, Dec 3 2004.

[187] R. Mohana-Borges, N. K. Goto, G. J. A. Kroon, H. J. Dyson, and P. E. Wright, "Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings," *Journal Of Molecular Biology,* vol. 340, pp. 1131-1142, Jul 23 2004.

[188] A. K. Jha, A. Colubri, K. F. Freed, and T. R. Sosnick, "Statistical coil model of the unfolded state: Resolving the reconciliation problem," *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 102, pp. 13099-13104, Sep 13 2005.

[189] H. J. Feldman and C. W. V. Hogue, "A fast method to sample real protein conformational space," *Proteins-Structure Function And Genetics,* vol. 39, pp. 112-131, May 1 2000.

[190] A. Marx, C. Nugoor, J. Muller, S. Panneerselvam, T. Timm, M. Bilang, E. Mylonas, D. I. Svergun, E. M. Mandelkow, and E. Mandelkow, "Structural variations in the catalytic and ubiquitin-associated domains of microtubule-associated protein/microtubule affinity regulating kinase (MARK) 1 and MARK2," *Journal of Biological Chemistry,* vol. 281, pp. 27586-27599, Sep 15 2006.

[191] Y. Shen and A. Bax, "Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology," *Journal Of Biomolecular Nmr,* vol. 38, pp. 289-302, Aug 2007.

[192] X. P. Xu and D. A. Case, "Automated prediction of N-15, C-13(alpha), C-13(beta) and C-13 ' chemical shifts in proteins using a density functional database," *Journal of Biomolecular Nmr,* vol. 21, pp. 321-333, Dec 2001.

[193] J. Meiler, "PROSHIFT: Protein chemical shift prediction using artificial neural networks," *Journal Of Biomolecular Nmr,* vol. 26, pp. 25-37, May 2003.

[194] D. S. Wishart, B. D. Sykes, and F. M. Richards, "Relationship between Nuclear-Magnetic-Resonance Chemical-Shift and Protein Secondary Structure," *Journal of Molecular Biology,* vol. 222, pp. 311-333, Nov 20 1991.

[195] G. Cornilescu, F. Delaglio, and A. Bax, "Protein backbone angle restraints from searching a database for chemical shift and sequence homology," *Journal of Biomolecular Nmr,* vol. 13, pp. 289-302, Mar 1999.

[196] J. E. Kohn, I. S. Millett, J. Jacob, B. Zagrovic, T. M. Dillon, N. Cingel, R. S. Dothager, S. Seifert, P. Thiyagarajan, T. R. Sosnick, M. Z. Hasan, V. S. Pande, I. Ruczinski, S. Doniach, and K. W. Plaxco, "Random-coil behavior and the dimensions of chemically unfolded proteins," *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 101, pp. 12491-12496, Aug 24 2004.

[197] A. Cavalli, X. Salvatella, C. M. Dobson, and M. Vendruscolo, "Protein structure determination from NMR chemical shifts," *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 104, pp. 9615-9620, Jun 2007.

[198] H. P. Gong, Y. Shen, and G. D. Rose, "Building native protein conformation from NMR backbone chemical shifts using Monte Carlo fragment assembly," *Protein Science,* vol. 16, pp. 1515-1521, Aug 2007.

[199] I. Khlistunova, M. Pickhardt, J. Biernat, Y. P. Wang, E. M. Mandelkow, and E. Mandelkow, "Inhibition of tau aggregation in cell models of tauopathy," *Current Alzheimer Research,* vol. 4, pp. 544-546, 2007.

[200] L. C. Serpell, J. Berriman, R. Jakes, M. Goedert, and R. A. Crowther, "Fiber diffraction of synthetic alpha-synuclein filaments shows amyloid-like cross-beta conformation," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 97, pp. 4897-4902, Apr 25 2000.

[201] H. Inouye, P. E. Fraser, and D. A. Kirschner, "Structure of Beta-Crystallite Assemblies Formed by Alzheimer Beta-Amyloid Protein Analogs - Analysis by X-Ray-Diffraction," *Biophysical Journal,* vol. 64, pp. 502-519, Feb 1993.

[202] K. Minoura, K. Tomoo, T. Ishida, H. Hasegawa, M. Sasaki, and T. Taniguchi, "Amphipathic helical behavior of the third repeat fragment in the tau microtubule-binding domain, studied by H-1 NMR spectroscopy," *Biochemical And Biophysical Research Communications,* vol. 294, pp. 210-214, Jun 7 2002.

[203] K. Minoura, T. M. Yao, K. Tomoo, M. Sumida, M. Sasaki, T. Taniguchi, and T. Ishida, "Different associational and conformational behaviors between the second and third repeat fragments in the tau microtubule-binding domain," *European Journal Of Biochemistry,* vol. 271, pp. 545-552, Feb 2004.

[204] M. Louhivuori, K. Paakkonen, K. Fredriksson, P. Permi, J. Lounila, and A. Annila, "On the origin of residual dipolar couplings from denatured proteins," *Journal Of The American Chemical Society,* vol. 125, pp. 15647-15650, Dec 17 2003.

[205] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization By Simulated Annealing," *Science,* vol. 220, pp. 671-680, 1983.

[206] C. M. Stultz and M. Karplus, "Dynamic ligand design and combinatorial optimization: Designing inhibitors to endothiapepsin," *Proteins-Structure Function And Genetics,* vol. 40, pp. 258-289, Aug 1 2000.

[207] J. D. Nulton and P. Salamon, "Statistical-Mechanics Of Combinatorial Optimization," *Physical Review A,* vol. 37, pp. 1351-1356, Feb 15 1988.

[208] T. F. Coleman and Y. Y. Li, "A reflective Newton method for minimizing a quadratic function subject to bounds on some of the variables," *Siam Journal On Optimization,* vol. 6, pp. 1040-1058, Nov 1996.

[209] T. F. Coleman and Y. Y. Li, "An interior trust region approach for nonlinear minimization subject to bounds," *Siam Journal On Optimization,* vol. 6, pp. 418-445, May 1996.

[210] B. K. Ho and K. A. Dill, "Folding very short peptides using molecular dynamics," *PLoS Comput Biol,* vol. 2, p. e27, Apr 2006.

[211] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson, "Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation," *J Mol Biol,* vol. 285, pp. 1735-47, Jan 29 1999.

[212] M. Feig, J. Karanicolas, and C. L. Brooks, 3rd, "MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology," *J Mol Graph Model,* vol. 22, pp. 377-95, May 2004.

[213] I. Radhakrishnan, G. C. Perez-Alvarado, H. J. Dyson, and P. E. Wright, "Conformational preferences in the Ser133-phosphorylated and non-phosphorylated forms of the kinase inducible transactivation domain of CREB," *FEBS Lett,* vol. 430, pp. 317-22, Jul 3 1998.

[214] R. Dawson, L. Muller, A. Dehner, C. Klein, H. Kessler, and J. Buchner, "The N-terminal domain of p53 is natively unfolded," *J Mol Biol,* vol. 332, pp. 1131-41, Oct 3 2003.

[215] A. S. Kim, L. T. Kakalis, N. Abdul-Manan, G. A. Liu, and M. K. Rosen, "Autoinhibition and activation mechanisms of the Wiskott-Aldrich syndrome protein," *Nature,* vol. 404, pp. 151-8, Mar 9 2000.

[216] E. R. Lacy, I. Filippov, W. S. Lewis, S. Otieno, L. Xiao, S. Weiss, L. Hengst, and R. W. Kriwacki, "p27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding," *Nat Struct Mol Biol,* vol. 11, pp. 358-64, Apr 2004.

[217] E. A. Bienkiewicz, J. N. Adkins, and K. J. Lumb, "Functional consequences of preorganized helical structure in the intrinsically disordered cell-cycle inhibitor p27(Kip1)," *Biochemistry,* vol. 41, pp. 752-9, Jan 22 2002.

[218] P. Tompa, "The interplay between structure and function in intrinsically unstructured proteins," *FEBS Lett,* vol. 579, pp. 3346-54, Jun 13 2005.

[219] M. Fuxreiter, I. Simon, P. Friedrich, and P. Tompa, "Preformed structural elements feature in partner recognition by intrinsically unstructured proteins," *J Mol Biol,* vol. 338, pp. 1015-26, May 14 2004.

[220] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic, "Intrinsically disordered protein," *J Mol Graph Model,* vol. 19, pp. 26-59, 2001.

[221] P. E. Wright and H. J. Dyson, "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm," *J Mol Biol,* vol. 293, pp. 321-31, Oct 22 1999.

[222] R. W. Kriwacki, L. Hengst, L. Tennant, S. I. Reed, and P. E. Wright, "Structural studies of p21(Waf1/Cip1/Sdi1) in the free and Cdk2-bound state: Conformational disorder mediates binding diversity," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 93, pp. 11504-11509, Oct 1996.

[223] J. W. Harper, G. R. Adami, N. Wei, K. Keyomarsi, and S. J. Elledge, "The p21 Cdk-interacting protein Cip1 is a potent inhibitor of G1 cyclin-dependent kinases," *Cell,* vol. 75, pp. 805-16, Nov 19 1993.

[224] S. Waga, G. J. Hannon, D. Beach, and B. Stillman, "The p21 inhibitor of cyclin-dependent kinases controls DNA replication by interaction with PCNA," *Nature,* vol. 369, pp. 574-8, Jun 16 1994.

[225] B. Levkau, H. Koyama, E. W. Raines, B. E. Clurman, B. Herren, K. Orth, J. M. Roberts, and R. Ross, "Cleavage of p21Cip1/Waf1 and p27Kip1 mediates apoptosis in endothelial cells through activation of Cdk2: role of a caspase cascade," *Mol Cell,* vol. 1, pp. 553-63, Mar 1998.

[226] G. P. Dotto, "p21(WAF1/Cip1): more than a break to the cell cycle?," *Biochim Biophys Acta,* vol. 1471, pp. M43-56, Jul 31 2000.

[227] L. Delavaine and N. B. La Thangue, "Control of E2F activity by p21Waf1/Cip1," *Oncogene,* vol. 18, pp. 5381-92, Sep 23 1999.

[228] J. O. Funk, S. Waga, J. B. Harry, E. Espling, B. Stillman, and D. A. Galloway, "Inhibition of CDK activity and PCNA-dependent DNA replication by p21 is blocked by interaction with the HPV-16 E7 oncoprotein," *Genes Dev,* vol. 11, pp. 2090-100, Aug 15 1997.

[229] T. Cheng, N. Rodrigues, H. Shen, Y. Yang, D. Dombkowski, M. Sykes, and D. T. Scadden, "Hematopoietic stem cell quiescence maintained by p21cip1/waf1," *Science,* vol. 287, pp. 1804-8, Mar 10 2000.

[230] R. Li, S. Waga, G. J. Hannon, D. Beach, and B. Stillman, "Differential effects by the p21 CDK inhibitor on PCNA-dependent DNA replication and repair," *Nature,* vol. 371, pp. 534-7, Oct 6 1994.

[231] J. Chen, P. K. Jackson, M. W. Kirschner, and A. Dutta, "Separate domains of p21 involved in the inhibition of Cdk kinase and PCNA," *Nature,* vol. 374, pp. 386-8, Mar 23 1995.

[232] M. Taules, A. Rodriguez-Vilarrupla, E. Rius, J. M. Estanyol, O. Casanovas, D. B. Sacks, E. Perez-Paya, O. Bachs, and N. Agell, "Calmodulin binds to p21(Cip1) and is involved in the regulation of its nuclear localization," *J Biol Chem,* vol. 274, pp. 24445-8, Aug 27 1999.

[233] J. M. Estanyol, M. Jaumot, O. Casanovas, A. Rodriguez-Vilarrupla, N. Agell, and O. Bachs, "The protein SET regulates the inhibitory effect of p21(Cip1) on cyclin E-cyclin-dependent kinase 2 activity," *J Biol Chem,* vol. 274, pp. 33161-5, Nov 12 1999.

[234] H. Kitaura, M. Shinshi, Y. Uchikoshi, T. Ono, S. M. Iguchi-Ariga, and H. Ariga, "Reciprocal regulation via protein-protein interaction between c-Myc and p21(cip1/waf1/sdi1) in DNA replication and transcription," *J Biol Chem,* vol. 275, pp. 10477-83, Apr 7 2000.

[235] V. Esteve, N. Canela, A. Rodriguez-Vilarrupla, R. Aligue, N. Agell, I. Mingarro, O. Bachs, and E. Perez-Paya, "The structural plasticity of the C terminus of p21Cip1 is a determinant for target protein recognition," *Chembiochem,* vol. 4, pp. 863-9, Sep 5 2003.

[236] K. L. Yap, J. Kim, K. Truong, M. Sherman, T. Yuan, and M. Ikura, "Calmodulin target database," *J Struct Funct Genomics,* vol. 1, pp. 8-14, 2000.

[237] D. S. Wishart and B. D. Sykes, "Chemical shifts as a tool for structure determination," *Methods Enzymol,* vol. 239, pp. 363-92, 1994.

[238] O. Zhang and J. D. Forman-Kay, "NMR studies of unfolded states of an SH3 domain in aqueous solution and denaturing conditions," *Biochemistry,* vol. 36, pp. 3959-70, Apr 1 1997.

[239] J. Yao, J. Chung, D. Eliezer, P. E. Wright, and H. J. Dyson, "NMR structural and dynamic characterization of the acid-unfolded state of apomyoglobin provides insights into the early events in protein folding," *Biochemistry,* vol. 40, pp. 3561-71, Mar 27 2001.

[240] J. A. Williamson and A. D. Miranker, "Direct detection of transient alpha-helical states in islet amyloid polypeptide," *Protein Sci,* vol. 16, pp. 110-7, Jan 2007.

[241] T. K. Mal, N. R. Skrynnikov, K. L. Yap, L. E. Kay, and M. Ikura, "Detecting protein kinase recognition modes of calmodulin by residual dipolar couplings in solution NMR," *Biochemistry,* vol. 41, pp. 12899-906, Oct 29 2002.

[242] G. M. Contessa, M. Orsale, S. Melino, V. Torre, M. Paci, A. Desideri, and D. O. Cicero, "Structure of calmodulin complexed with an olfactory CNG channel fragment and role of the central linker: residual dipolar couplings to evaluate calmodulin binding modes outside the kinase family," *J Biomol NMR,* vol. 31, pp. 185-99, Mar 2005.

[243] J. A. Losonczi, M. Andrec, M. W. Fischer, and J. H. Prestegard, "Order matrix analysis of residual dipolar couplings using singular value decomposition," *J Magn Reson,* vol. 138, pp. 334-42, Jun 1999.

[244] W. E. Meador, A. R. Means, and F. A. Quiocho, "Modulation of calmodulin plasticity in molecular recognition on the basis of x-ray structures," *Science,* vol. 262, pp. 1718-21, Dec 10 1993.

[245] N. Hayashi, M. Matsubara, A. Takasaki, K. Titani, and H. Taniguchi, "An expression system of rat calmodulin using T7 phage promoter in Escherichia coli," *Protein Expr Purif,* vol. 12, pp. 25-8, Feb 1998.

[246] N. Tjandra and A. Bax, "Large Variations in 13C Chemical Shift Anisotropy in Proteins Correlate with Secondary Structure," *J. Am. Chem. Soc.,* vol. 119, pp. 9576 - 9577, 1997.

[247] G. W. Vuister and A. Bax, "Quantitative J correlation: a new approach for measuring homonuclear three-bond J(HNH.alpha.) coupling constants in 15N-enriched proteins," *J. Am. Chem. Soc.,* vol. 115, pp. 7772-7777, 1993.

[248] D. S. Wishart, C. G. Bigam, J. Yao, F. Abildgaard, H. J. Dyson, E. Oldfield, J. L. Markley, and B. D. Sykes, "1H, 13C and 15N chemical shift referencing in biomolecular NMR," *J Biomol NMR,* vol. 6, pp. 135-40, Sep 1995.

[249] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, and A. Bax, "NMRPipe: a multidimensional spectral processing system based on UNIX pipes," *J Biomol NMR,* vol. 6, pp. 277-93, Nov 1995.

[250] R. Keller, *The computer aided resonance assignment tutorial*. Goldau, Switzerland, 2004.

[251] S. Schwarzinger, G. J. Kroon, T. R. Foss, P. E. Wright, and H. J. Dyson, "Random coil chemical shifts in acidic 8 M urea: implementation of random coil shift data in NMRView," *J Biomol NMR,* vol. 18, pp. 43-8, Sep 2000.

[252] L. J. Smith, K. A. Bolin, H. Schwalbe, M. W. MacArthur, J. M. Thornton, and C. M. Dobson, "Analysis of main chain torsion angles in proteins: prediction of NMR coupling constants for native and random coil conformations," *J Mol Biol,* vol. 255, pp. 494-506, Jan 26 1996.

[253] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research,* vol. 28, pp. 235-242, 2000.